

MULTIPLE PENALIZED PRINCIPAL CURVES: ANALYSIS AND COMPUTATION

SLAV KIROV AND DEJAN SLEPČEV

ABSTRACT. We study the problem of determining the one-dimensional structure that best represents a given data set. More precisely, we take a variational approach to approximating a given measure (data) by curves. We consider an objective functional whose minimizers are a regularization of principal curves and introduce a new functional which allows for multiple curves. We prove existence of minimizers and investigate their properties. While both of the functionals used are non-convex, we show that enlarging the configuration space to allow for multiple curves leads to a simpler energy landscape with fewer undesirable (high-energy) local minima. We provide an efficient algorithm for approximating minimizers of the functional and demonstrate its performance on real and synthetic data. The numerical examples illustrate the effectiveness of the proposed approach in the presence of substantial noise, and the viability of the algorithm for high-dimensional data.

Keywords. principal curves, geometry of data, curve fitting

Classification. 49M25, 65D10, 62G99, 65D18, 65K10, 49Q20

1. INTRODUCTION

We consider the problem of finding one-dimensional structures best representing the data given as point clouds. This is a classical problem. It has been studied in 80's by Hastie and Stuetzle [23] who introduced *principal curves* as the curves going through the “middle” of the data. A number of modifications of principal curves, which make them more stable and easier to compute, followed [12, 15, 21, 25, 36]. However there are very few precise mathematical results on the relation between the properties of principal curves and their variants, and the geometry of the data. On the other hand rigorous mathematical setup has been developed for related problems studied in the context of optimal (transportation) network design [5, 6]. In particular, an objective functional studied in network design, the average-distance functional, is closely related to a regularization of principal curves.

Our first aim is to carefully investigate a desirable variant of principal curves suggested by the average-distance problem. The objective functional includes a length penalty for regularization, and we call its minimizers *penalized principal curves*. We establish their basic properties (existence and basic regularity of minimizers) and investigate the sense in which they approximate the data and represent the one-dimensional structure. One of the shortcomings of principal curves is that they tend to overfit noisy data. Adding a regularization term to the objective functional minimized by principal curves is a common way to address overfitting. The drawback is that doing so introduces bias: when data lie on a smooth curve the minimizer is only going to approximate them. We investigate the relationship between the data and the minimizers and establish how the length scales present in the data and the parameters of the functional dictate the length scales seen in the minimizers. In particular we provide the critical length scale below which variations in the input data are treated as noise and establish the typical error (bias) when the input curve is smooth. We emphasize that the former has direct implications for when penalized principal curves begin to overfit given data.

Our second aim is to introduce a functional that permits its minimizers to consist of more than one curve (*multiple penalized principal curves*). The motivation is twofold. The data itself may have one-dimensional structure that consists of more than one component, and the relaxed setting would allow it to be appropriately represented. The less immediate appeal of the new functional is that it guides the design of an improved scheme for computing penalized principal curves. Namely, for many datasets the penalized principal curves functional has a complicated energy landscape with many local minima. This is a typical situation and an issue for virtually all present approaches to nonlinear principal components. As we explain below, enlarging the set over which the functional is considered (from a single curve to multiple curves) and appropriately penalizing the number of components leads to significantly better behavior of energy descent methods (they more often converge to low-energy local minima).

We find topological changes of multiple penalized principal curves are governed by a critical *linear density*. The linear density of a curve is the density of the projected data on the curve with respect to its length. If the linear density of a single curve drops below the critical value over a large enough length scale, a lower-energy configuration consisting of two curves can be obtained by removing the corresponding curve segment. Such steps are the means by which configurations following energy descent stay in higher-density regions of the data, and avoid local minima that penalized principal curves are vulnerable to. Identification of the critical linear density and the length scale over which it is recognized by the functional further provide insight as to the conditions under and the resolution to which minimizers can recover one-dimensional components of the data.

We apply modern optimization algorithms based on alternating direction method of multipliers (ADMM) [3] and closely related Bregman iterations [22, 30] to compute approximate minimizers. We describe the algorithm in detail, discuss its complexity and present computational examples that both illustrate the theoretical findings and support the viability of the approach for finding complex one-dimensional structures in point clouds with tens of thousands of points in high dimensions.

1.1. Outline. In Section 2 we introduce the objective functionals (both for single and multiple curve approximation), and recall some of the related approaches. We establish basic properties of the functionals, including the existence of minimizers and their regularity. Under assumption of smoothness we derive the Euler-Lagrange equation for critical points of the functional. We conclude Section 2 by computing the second variation of the functional. In Section 3 we provide a number of illustrative examples and investigate the relation between the length scales present in the data, the parameters of the functional and the length scales present in the minimizers. At the end of Section 3 we discuss parameter selection for the functional. In Section 4 we describe the algorithm for computing approximate minimizers of the (MPPC) functional. In Section 4.7 we provide some further numerical examples that illustrate the applicability of the functionals and algorithm. Section 5 contains the conclusion and a brief discussion of and comparison with other approaches for one-dimensional data cloud approximation. Appendix A contains some technical details of an analysis of a minimizer considered in Section 3.

2. THE FUNCTIONALS AND BASIC PROPERTIES

In this section we introduce the *penalized principal curves functional* and the *multiple penalized principal curves functional*. We recall and prove some of their basic properties. Let \mathcal{M} be the set of finite, compactly supported measures on \mathbb{R}^d , with $d \geq 2$ and $\mu(\mathbb{R}^d) > 0$.

2.1. Penalized principal curves. Given a measure (distribution of data) $\mu \in \mathcal{M}$, $\lambda > 0$, and $p \geq 1$, the *penalized principal curves* are minimizers of

$$(PPC) \quad E_\mu^\lambda(\gamma) := \int_{\mathbb{R}^d} d(x, \Gamma)^p d\mu(x) + \lambda L(\gamma)$$

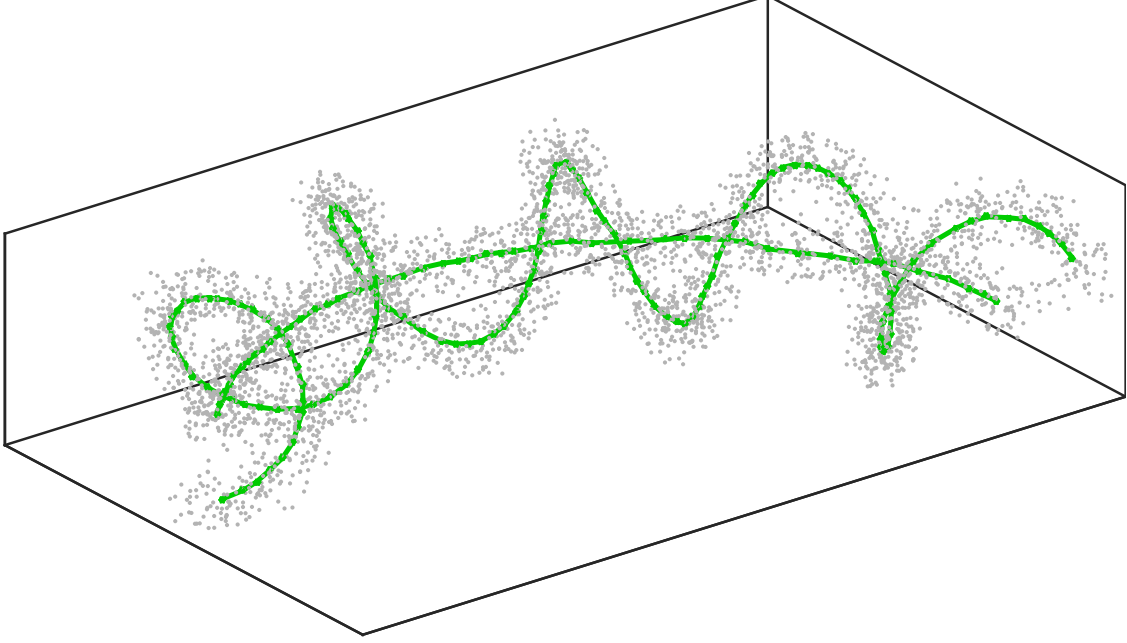


FIGURE 1. Example of a point cloud generated by noisy samples of two curves (not shown): a section of a circle and a curved helix wrapping around it. The green curves shown represent the one dimensional approximation of the data cloud obtained by minimizing the proposed functional (MPPC) using the algorithm of Section 4.

over $\gamma \in \mathcal{C} := \{\gamma : [0, a] \rightarrow \mathbb{R}^d : a \geq 0, \gamma \text{ is Lipschitz with } |\gamma'| \leq 1, \mathcal{L}^1 - \text{a.e.}\}$, and where $\Gamma := \gamma([0, a])$, $d(x, \Gamma)$ is the distance from x to set Γ and $L(\gamma)$ is the length of γ :

$$L(\gamma) := \|\gamma\|_{TV} := \sup \left\{ \sum_{i=2}^n |\gamma(x_i) - \gamma(x_{i-1})| : 0 \leq x_1 < x_2 < \dots < x_n \leq a, n \in \mathbb{N} \right\}$$

and where $|\cdot|$ denotes the Euclidean norm. The functional is closely related to the average-distance problem introduced by Buttazzo, Oudet, and Stepanov [5] having in mind applications to optimal transportation networks [6]. In this context, the first term can be viewed as the cost of a population to reach the network, and the second a cost of the network itself. There the authors considered general connected one-dimensional sets and instead of length penalty considered a length constraint. The penalized functional for one-dimensional sets was studied by Lu and one of the authors in [28], and later for curves (as in this paper) in [27]. Similar functionals have been considered in statistics and machine learning literature as regularizations of the principal curves problem by Tibshirani [37] (introduces curvature penalization) Kegl, Krzyzak, Linder, and Zeger [25] (length constraint), Biau and Fischer [2] (length constraint) Smola, Mika, Schölkopf, and Williamson [36] (a variety of penalizations including penalizing length as in (PPC)) and others.

The first term in (PPC) measures the approximation error, while the second one penalizes the complexity of the approximation. If μ has smooth density, or is concentrated on a smooth curve, the minimizer γ is typically supported on smooth curves. However this is not universally true. Namely it was shown in [35] that minimizers of average-distance problems can have corners, even if μ has smooth density. An analogous argument applies to (PPC) (and later introduced (MPPC)). This raises important modeling questions regarding what the best functional is and if further regularization is appropriate. We do not address these questions in this paper.

Existence of minimizers of (PPC) in \mathcal{C} was shown in [27]. There it was also shown that any minimizer γ_{\min} has the following total curvature bound

$$\|\gamma'_{\min}\|_{TV} \leq \frac{p}{\lambda} \text{diam}(\text{supp}(\mu))^{p-1} \mu(\mathbb{R}^d).$$

The total variation (TV) above allows to treat the curvature as a measure, with delta masses at locations of corners, which is necessary in light of the possible lack of regularity. In [27], it was also shown that minimizing curves are injective (i.e. do not self-intersect) in dimension $d = 2$ if $p \geq 2$.

2.2. Multiple penalized principal curves. We now introduce an extension of (PPC) which allows for configurations to consist of more than one component. Since (PPC) can be made arbitrarily small by considering γ with many components, a penalty on the number of components is needed. Thus we propose the following functional for multiple curves

$$(MPPC) \quad E_{\mu}^{\lambda_1, \lambda_2}(\gamma) := \int_{\mathbb{R}^d} d(x, \Gamma)^p d\mu(x) + \lambda_1 (L(\gamma) + \lambda_2 (k(\gamma) - 1))$$

where, we relax γ to now be piecewise Lipschitz and $k(\gamma)$ is the number of curves used to parametrize γ . More precisely, we aim to minimize (MPPC) over the admissible set

$$\mathcal{A} := \left\{ \gamma = \{\gamma^i\}_{i=1}^{k(\gamma)} : k(\gamma) \in \mathbb{N}, \gamma^i \in \mathcal{C}, i = 1, \dots, k(\gamma) \right\}.$$

We call elements of \mathcal{A} *multiple curves*. One may think of this functional as penalizing both zero- and one-dimensional complexities of approximations to μ . In particular we can recover the (PPC) functional by taking λ_2 large enough. On the other hand, taking λ_1 large enough leads to a k-means clustering problem which penalizes the number of clusters, and has been encountered in [4, 26].

The main motivation for considering (MPPC), even if only one curve is sought, has to do with the non-convexity of the (PPC). We will see that numerically minimizing (MPPC) often helps evade undesirable (high-energy) local minima of the (PPC) functional. In particular (MPPC) can be seen as a relaxation of (PPC) to a larger configuration space. The energy descent for (MPPC) allows for curve splitting and reconnecting which is the mechanism that enables one to evade local minima of (PPC).

2.3. Existence of minimizers of (MPPC). We show that minimizers of (MPPC) exist in \mathcal{A} . We follow the approach of [27], where existence of minimizers was shown for (PPC). We first cover some preliminaries, including defining the distance between curves. If $\gamma_1, \gamma_2 \in \mathcal{C}$ with respective domains $[0, a_1], [0, a_2]$, where $a_1 \leq a_2$, we define the extension of γ_1 to $[0, a_2]$ as

$$\tilde{\gamma}_1(t) = \begin{cases} \gamma_1(t) & \text{if } t \in [0, a_1] \\ \gamma_1(a_1) & \text{if } t \in (a_1, a_2]. \end{cases}$$

We let

$$d_{\mathcal{C}}(\gamma_1, \gamma_2) = \max_{t \in [0, a_2]} |\tilde{\gamma}_1(t) - \gamma_2(t)|.$$

We have the following lemma, and the subsequent existence of minimizers.

Lemma 2.1. *Consider a measure $\mu \in \mathcal{M}$ and $\lambda_1, \lambda_2 > 0, p \geq 1$.*

- (i) *For any minimizing sequence $\{\gamma_n\}$ of (MPPC)*
 - (a) $\limsup_{n \rightarrow \infty} k(\gamma_n) \leq \frac{1}{\lambda_1 \lambda_2} (\text{diam supp}(\mu))^p$, and
 - (b) $\limsup_{n \rightarrow \infty} L(\gamma_n) \leq \frac{1}{\lambda_1} (\text{diam supp}(\mu))^p$
- (ii) *There exists a minimizing sequence $\{\gamma_n\}$ of (MPPC) such that $\forall n, \Gamma_n$ is contained in $\text{Conv}(\mu)$, the convex hull of the support of μ .*

Proof. The first property follows by taking a singleton as a competitor. The second follows from projecting any minimizing sequence onto $\text{Conv}(\mu)$. Doing so can only decrease the energy, as shown in [6, 27]. The argument relies on the fact that projection onto a convex set decrease length. \square

Lemma 2.2. *Given a positive measure $\mu \in \mathcal{M}$ and $\lambda_1, \lambda_2 > 0, p \geq 1$, the functional (MPPC) has a minimizer in \mathcal{A} . Moreover, the image of any minimizer is contained in the convex hull of the support of μ .*

Proof. The proof is an extension of the one found in [27] for (PPC). Let $\{\gamma_n\}_{n \in \mathbb{N}}$ be a minimizing sequence in \mathcal{A} . Since the number of curves $k(\gamma_n)$ is bounded, we can find a subsequence (which we take to be the whole sequence) with each member having the same number of curves k . We enumerate the curves in each member of the sequence as $\gamma_n = \{\gamma_n^i\}_{i=1}^k$. We assume that each curve γ_n^i is arc-length parametrized for all $n \in \mathbb{N}, i \leq k$. Since the lengths of the curves are uniformly bounded, let $L = \sup_{n,i} L(\gamma_n^i)$, and extend the parametrization for each curve in the way defined above. Then for each $i \leq k$, the curves $\{\gamma_n^i\}_{n \in \mathbb{N}}$ satisfy the hypotheses of the Arzela-Ascoli Theorem. Hence for each $i \leq k$, up to a subsequence γ_n^i converge uniformly to a curve $\gamma^i : [0, L] \rightarrow \mathbb{R}^d$. Diagonalizing, we find a subsequence (which we take to be the whole sequence) for which the aforementioned convergence holds for all $i \leq k$. Moreover, the limiting object is a collection of curves which are 1-Lipschitz since all of the curves in the sequence are. Thus $\gamma := \{\gamma^i\}_{i=1}^k \in \mathcal{A}$.

The mapping $\Gamma \mapsto \int_{\mathbb{R}^d} d(x, \Gamma)^p d\mu(x)$ is continuous and $\Gamma \mapsto L(\Gamma)$ is lower-semicontinuous with respect to convergence in \mathcal{C} . Thus $\liminf_{n \rightarrow \infty} E_{\mu}^{\lambda_1, \lambda_2, p}(\gamma_n) \geq E_{\mu}^{\lambda_1, \lambda_2, p}(\gamma)$, and so γ is a minimizer. \square

2.4. First variation. In this section, we compute the first interior variation of the (PPC) functional and state under what conditions a smooth curve is a critical (i.e. stationary) point. In the case of multiple curves, one can apply the following analysis to each curve separately.

Let μ be a compactly supported measure. We will assume the curve $\gamma : [a, b] \rightarrow \mathbb{R}^d$ is C^2 . Although minimizers may have corners as mentioned earlier, our use of the first variation is aimed at understanding how the parameters of the functional relate to the length scales observed in minimizers. We generally expect that minimizers will be C^2 except at finitely many points, and hence that our analysis applies to intervals between such points.

To compute the first variation we only perturb the interior of the curve, and not the endpoints. Without loss of generality we assume that $|\gamma_s| = 1$, where γ_s denotes the partial derivative in s . We consider variations of γ of the form $\gamma(s, t) = \gamma(s) + tv(s)$, where $v \in C^2([a, b], \mathbb{R}^d)$, $v(a) = v(b) = 0$. We note that one could allow for $v(a)$ and $v(b)$ to be nonzero (as has been considered for example in [35]), but it is not needed for our analysis. We can furthermore assume (by reparameterizing the curves if necessary) that v is orthogonal to the curve: $v(s) \cdot \gamma_s(s) = 0 \quad \forall s \in [a, b]$.

Let $\Gamma_t := \gamma([a, b], t)$, the image of $\gamma(\cdot, t)$. To compute how the energy (PPC) is changing when γ is perturbed we need to describe how the distance of points to Γ_t is changing with t . For any $x \in \text{supp } \mu$ let $\Pi_t(x)$ be a point on Γ_t which is closest to x , that is let $\Pi_t(x)$ be a minimizer of $|x - y|$ over $y \in \Gamma_t$. For simplicity, we assume that $\Pi_t(x)$ is unique for all $x \in \text{supp } \mu$. The general case that the closest point is non-unique can also be considered [27, 35], but we omit it here. A further reason this assumption is inconsequential is that since μ is absolutely continuous with respect to the Lebesgue measure \mathcal{L}^d , the set of points where Π_t is non-unique has μ -measure zero [29]. We call Π_t the projection of data onto Γ_t . For $x \in \text{supp } \mu$ let $g(x, t) := d(x, \Gamma_t)^2 = |x - \Pi_t(x)|^2$. Then

$$(2.1) \quad \begin{aligned} \frac{\partial g}{\partial t} &= -2(x - \Pi_t(x)) \cdot \gamma_t \\ \frac{\partial^2 g}{\partial t^2} &= 2 \left(|\gamma_t|^2 - (x - \Pi_t(x)) \cdot \gamma_{tt} - \frac{(\gamma_t \cdot \gamma_s - (x - \Pi_t(x)) \cdot \gamma_{st})^2}{|\gamma_s|^2 - (x - \Pi_t(x)) \cdot \gamma_{ss}} \right) \end{aligned}$$

where γ and its derivatives are evaluated at $(s(t), t)$, where $s = \gamma(\cdot, t)^{-1}(\Pi_t(x))$, that is $s(t) = \arg \min_{r \in [a, b]} d(x, \gamma(t, r))$. Note that for any $s \in (a, b)$ the set of points projecting onto $\gamma(s, t)$

satisfies $\Pi_t^{-1}(\gamma(s, t)) \subset \gamma_s(s, t)^\perp$. Taking the derivative in t , and changing coordinates so that the approximation-error term is written as double integral, we obtain

$$\frac{dE}{dt} = \int_a^b \left(\lambda_1 \frac{\gamma_s}{|\gamma_s|} \cdot \gamma_{st} - 2\alpha(s) \int_{\Pi_t^{-1}(\gamma)} (x - \Pi_t(x)) \cdot \gamma_t |1 - \vec{\mathcal{K}}(s) \cdot (x - \Pi_t(x))| d\mu_s(x) \right) ds$$

where we have suppressed notation for dependence on t (and in some places s), $|1 + \vec{\mathcal{K}}(s) \cdot (\gamma(s) - x)|$ is the Jacobian for change of coordinates, and $\vec{\mathcal{K}}$ is the curvature vector of γ . Here we have used the disintegration theorem (see for example pages 78-80 of [13]) to rewrite an integral for μ over \mathbb{R}^n as an iterated integral along slices orthogonal to the curve (more precisely over the set of points that project to a given point on the curve). We have denoted by α the linear density of the projection of μ to Γ_t , pulled back to the parameterization of γ ; that is $\alpha := d(\gamma(\cdot, t)^{-1} \circ \Pi_t)_\# \mu / d\mathcal{L}^1$. By μ_s we denote the probability measure supported on the slice $\Pi_t^{-1}(\gamma(s, t))$. Integrating by parts we obtain

$$\left. \frac{dE}{dt} \right|_{t=0} = \int_a^b \left(-\lambda_1 \vec{\mathcal{K}} \cdot \gamma_t - 2\alpha(s) \int_{\Pi_t^{-1}(\gamma)} (x - \Pi_t(x)) \cdot \gamma_t |1 - \vec{\mathcal{K}} \cdot (x - \Pi_t(x))| d\mu_s(x) \right) ds.$$

We conclude γ is a stationary configuration if and only if

$$(2.2) \quad \lambda_1 \vec{\mathcal{K}}(s) = -2\alpha(s) \int_{\Pi_t^{-1}(\gamma)} (x - \Pi_t(x)) |1 - \vec{\mathcal{K}}(s) \cdot (x - \Pi_t(x))| d\mu_s(x)$$

for \mathcal{L}^1 - a.e. $s \in (a, b)$.

2.5. Second variation. In this section we compute the second variation of (PPC) for the purpose of providing conditions for linear stability. That is, we focus on the case that a straight line segment is a stationary configuration (critical point), and find when it is stable under the considered perturbations (when the second variation is greater than zero). This has important implications for determining when the penalized principal curves start to overfit the data, and is further investigated in the next section.

If γ is a straight line segment, $\vec{\mathcal{K}} = 0$, and (2.2) simplifies to

$$\gamma(s) = \bar{x}(s) := \int_{\Pi_t^{-1}(\gamma(s, t))} x d\mu_s(x)$$

for \mathcal{L}^1 - a.e. $s \in (a, b)$ such that $\alpha(s) \neq 0$. This simply states that a straight line is a critical point of the functional if and only if almost every point on the line it is the mean of points projecting there. In other words, the condition is equivalent to γ being a principal curve (in the original sense).

The second variation of the length term is

$$(2.3) \quad \begin{aligned} \frac{d^2}{dt^2} L(\gamma) &= \int_a^b \left(\frac{\gamma_{st}}{|\gamma_s|} - \frac{\gamma_s}{|\gamma_s|^2} \left(\frac{\gamma_s}{|\gamma_s|} \cdot \gamma_{st} \right) \right) \cdot \gamma_{st} + \frac{\gamma_s}{|\gamma_s|} \cdot \gamma_{stt} ds \\ &= \int_a^b \frac{1}{|\gamma_s|} \left(|\gamma_{st}|^2 - \left(\frac{\gamma_s}{|\gamma_s|} \cdot \gamma_{st} \right)^2 \right) + \frac{\gamma_s}{|\gamma_s|} \cdot \gamma_{stt} ds \end{aligned}$$

We note that $0 = (\gamma_s \cdot \gamma_t)_s = \gamma_{ss} \cdot \gamma_t + \gamma_s \cdot \gamma_{st}$, and therefore $\gamma_s \cdot \gamma_{st} = 0$, so that the second variation of the length term becomes just $|\gamma_{st}|^2$. Thus using (2.1) we get

$$(2.4) \quad \left. \frac{d^2 E}{dt^2} \right|_{t=0} = \int_a^b \left(\lambda_1 |\gamma_{st}|^2 + 2\alpha(s) \int_{\Pi_t^{-1}(\gamma)} (|x - \Pi_t(x)|^2 - ((x - \Pi_t(x)) \cdot \gamma_{st})^2) d\mu_s(x) \right) ds.$$

3. RELATION BETWEEN THE MINIMIZERS AND THE DATA

In this section, our goal is to relate the parameters of the functional, the length-scales present in the data, and the length-scales seen in the minimizers. To do so we consider examples of data and corresponding minimizers, use the characterization of critical points of (MPPC), and perform linear stability analysis.

3.1. Examples and properties of minimizers. Here we provide some insight as to how minimizers of (MPPC) behave. We start by characterizing minimizers in some simple yet instructive cases. In the first couple of cases we focus on the behavior of single curves, and then investigate when minimizers develop multiple components.

3.1.1. Data on a curve. Here we study the bias of penalized principal curves when the data lie on a curve without noise. If μ is supported on the image of a smooth curve, and a local minimizer γ of (PPC) is sufficiently close to μ , one can obtain an exact expression for the projection distance. More precisely, suppose that for each $s \in (a, b)$, only one point in $\text{supp } \mu$ projects to it. That is $\forall s \in (a, b)$, the set of $x_s \in \text{supp } \mu$ such that s minimizes $|x_s - \gamma(\hat{s})|$ over $\hat{s} \in [a, b]$ is a singleton. Then (2.2) simplifies to

$$\lambda_1 \mathcal{K}(s) = 2\alpha(s)h(1 + \mathcal{K}(s)h)$$

where $h := |\gamma(s) - x_s|$, \mathcal{K} denotes the unsigned scalar curvature of γ , and α is the projected linear density. Consequently

$$(3.1) \quad h = \frac{1}{2\mathcal{K}} \left(\sqrt{1 + 2\frac{\lambda_1 \mathcal{K}^2}{\alpha}} - 1 \right) \approx \begin{cases} \sqrt{\frac{\lambda_1}{2\alpha}} & \text{if } \frac{1}{\mathcal{K}} \ll \sqrt{\frac{\lambda_1}{\alpha}} \\ \frac{\lambda_1 \mathcal{K}}{2\alpha} & \text{if } \frac{1}{\mathcal{K}} \gg \sqrt{\frac{\lambda_1}{\alpha}}. \end{cases}$$

Note that always $h \leq \sqrt{\frac{\lambda_1}{2\alpha}}$. We illustrate the transition of the projection distance l indicated in (3.1) with the example below.

Example 3.1. Curve with decaying oscillations. We consider data uniformly spaced on the image of the function $\frac{x}{5} \sin(-4\pi \log(x))$, which ensures that the amplitude and period are decreasing with the same rate, as $x \rightarrow 0^+$. In Figure 2, the linear density of the data is constant (with respect to arc length) with total mass 1, and solution curves are shown for two different values of λ_1 . For x small enough the minimizing curve is flat, as it is not influenced by oscillations whose amplitude is less than $\sqrt{\frac{\lambda_1}{\alpha}}$. As the amplitude of oscillations grows beyond the smoothing length scale the minimizing curves start to follow them. As x gets larger and \mathcal{K} becomes smaller, the projection distances at the peaks start to scale linearly with λ_1 , as predicted by (3.1). Indeed, as \mathcal{K} decreases to zero the ratio of the curvature of the minimizer to that of the data curve approaches one and α converges to a constant,. Hence from (3.1) follows that the ratio of the projection distances at the peaks converges to the ratio of the λ_1 values.

3.1.2. Linear stability. In this section we establish conditions for the linear stability of penalized principal curves. For simplicity we consider the case when $\text{supp } \mu \subset \mathbb{R}^2$. Suppose that $\gamma : [0, L] \rightarrow \mathbb{R}^2$ is arc-length parametrized and a stationary configuration of (PPC), and that for some $0 \leq a < b \leq L$, $\gamma([a, b])$ is a line segment. As previously, we let α denote the projected linear density of μ onto γ .

We evaluate the second variation (2.4) over the interval $[a, b]$, where the considered variations of γ are $\gamma_t(s) = v(s) = (v_1(s), v_2(s))$, where $\gamma_s \cdot \gamma_t = 0$. Since γ is a line segment on $[a, b]$, we can consider coordinates where $v_1(s) = 0$. We then have

$$\left. \frac{d^2 E}{dt^2} \right|_{t=0} = \int_a^b \lambda_1 (v_2')^2 + 2\alpha(s) \int_{\Pi_t^{-1}(\gamma)} \left(v_2^2 - (v_2'(x - \Pi_t(x)))^2 \right) d\mu_s(x) ds.$$

We define the *mean squared projection distance*

$$(3.2) \quad H(s) := \left(\int_{\Pi_t^{-1}(\gamma)} (x - \Pi_t(x))^2 d\mu_s(x) \right)^{\frac{1}{2}}$$

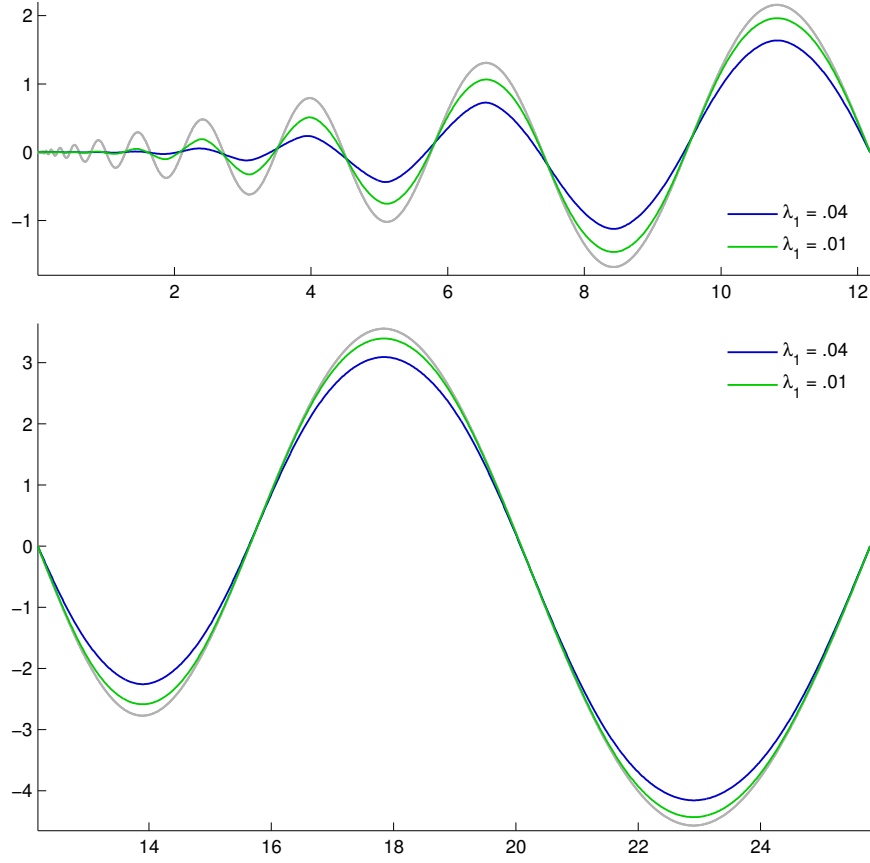


FIGURE 2. Numerical results shown for $n = 3000$ uniformly spaced data points (in gray) on the image of $\frac{x}{5} \sin(-4\pi \log(x))$ for $x \in [.001, e^{3.25}]$, and two different values of λ_1 .

and obtain

$$(3.3) \quad \left. \frac{d^2 E}{dt^2} \right|_{t=0} = \int_a^b (\lambda_1 - 2\alpha(s)H(s)^2) (v_2')^2 + 2\alpha(s)v_2^2 ds.$$

We see that if $\lambda_1 \geq 2\alpha(s)H(s)^2$ for almost every $s \in [a, b]$, then $\left. \frac{d^2 E}{dt^2} \right|_{t=0} > 0$ and so γ is linearly stable.

On the other hand, suppose that $\lambda_1 < 2\alpha(s)H(s)^2$ on some subinterval – without loss of generality we take it to be the entire interval $[a, b]$. Consider the perturbation given by $v_2(s) = \sin(ns)$. Then the RHS of (3.3) becomes

$$n^2 \int_a^b (\lambda_1 - 2\alpha(s)H(s)^2) \cos^2(ns) ds + 2 \int_a^b \alpha(s) \sin^2(ns) ds$$

and we see the first term dominates (in absolute value) the second for n large enough. Hence line segment

$$(3.4) \quad \gamma \text{ is linearly unstable on intervals where } \lambda_1 < 2\alpha H^2.$$

In the following examples we examine linear stability for some special cases of the data μ .

Example 3.2. Parallel lines. We start with a simple case in which data, μ , lie uniformly on two parallel lines. In Figure 3 we show computed local minimizers starting with a slight perturbation

of the initial straight line configuration, using the algorithm later described in Section 4. The data lines are of length 2, so that $\alpha = 0.5$ for the straight line configuration. The parameter $\lambda_1 = 0.16$ and hence the condition for linear instability (3.4) of the straight line steady state becomes $0.4 < H$. The numerical results show that indeed straight line steady state becomes unstable when H becomes slightly larger than 0.4.

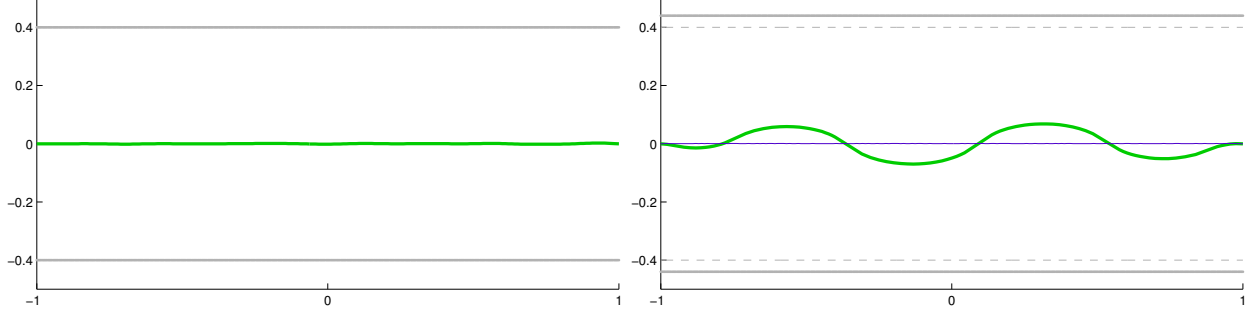


FIGURE 3. The data are gray line segments at height $H = \pm 0.4$ on the left image and $H = \pm 0.44$ on the right image. We numerically computed the local minimizers (green) of (MPPC) among curves with fixed endpoints at $(-1, 0)$ and $(1, 0)$, starting with slight perturbation of the line segment $[-1, 1] \times \{0\}$.

Example 3.3. Uniform density in rectangle. Consider a probability measure, μ , with uniform density over $[0, L] \times [0, 2h]$ with $L \gg h$. Linear instability of the line segment $\{\frac{1}{h}\} \times [0, L]$ (which is a critical point of (PPC)) can be seen as indication of when a local minimizer starts to overfit the data. It follows from (3.2) that $H^2 = \frac{1}{3}h^2$, and from (3.4) that $\lambda_1^* = \frac{2}{3L}h^2$ is the critical value for linear stability.

In Figure 4, we show the resulting local minimizers of (PPC) when starting from a small perturbation of the straight line, for several values of λ_1 , for $h = \frac{1}{2}$ and $L = 4$. The results from the numerical experiment appear to agree with the predicted critical value of $\lambda_1^* = 1/24$, as the computed minimizer corresponding to $\lambda_1 = 1/27$ has visible oscillations, while that of $\lambda_1 = 1/23$ does not.

To illustrate how closely the curves approximate that data we consider average mean projection distance, H , for various values of λ_1 . We expect that condition for linear stability (3.4), which was derived for straight-line critical points applies, approximately, to curved minimizers. In particular we expect that curves where H is larger than (approximately) $\sqrt{\frac{\lambda_1}{2\alpha}}$ will not be minimizers and will be evolved further by the algorithm. Here we investigate numerically if for minimizers $H \approx \sqrt{\frac{\lambda_1}{2\alpha}}$, as is the case in one regime of (3.1). Our findings are presented on Figure 5.

Example 3.4. Vertical Gaussian noise. Here we briefly remark on the case that μ has Gaussian noise with variance σ^2 orthogonal to a straight line. We note that the mean squared projection distance H is just the standard deviation σ . Therefore linear instability (overfitting) occurs if and only if $\lambda_1 < 2\alpha\sigma^2$.

3.1.3. Role of λ_2 . We now turn our attention to the role of λ_2 in (MPPC). Our goal is to understand when do transitions in the number of curves in minimizers occur.

By direct inspection of (MPPC), it is always energetically advantageous to connect endpoints of distinct curves if the distance between them is less than λ_2 . Similarly, it is never advantageous to disconnect a curve by removing a segment which has length less than λ_2 . Thus λ_2 represents

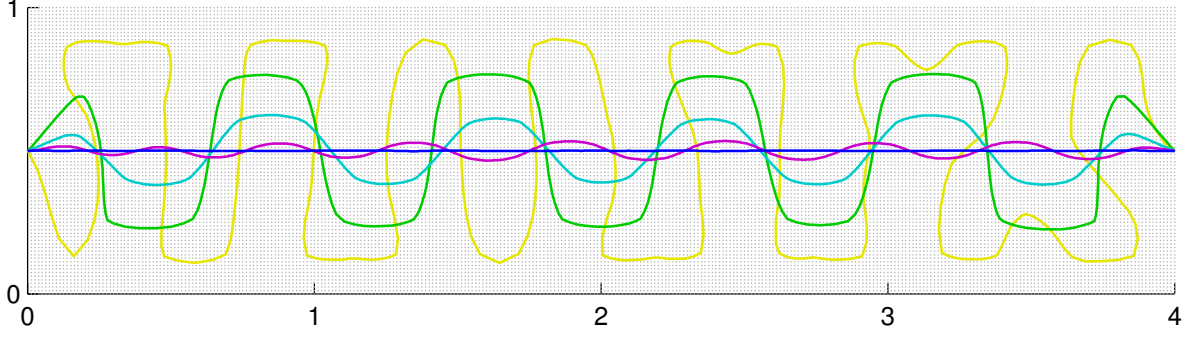


FIGURE 4. Numerical results showing local minimizers of (PPC) for various values of λ_1 . The data are a grid of $n = 361 \times 81$ uniformly spaced points with total mass equal to 1. Curves with decreasing amplitude correspond to $\lambda_1 = 1/1000, 1/150, 1/50, 1/27, 1/23$. Recall that the critical value for linear stability is $\lambda_1^* = 1/24$. The initial curve used for all results was a randomly perturbed straight line segment $[0, 4] \times \{\frac{1}{2}\}$. The endpoints were kept fixed at $(0, 0.5), (4, 0.5)$ to avoid boundary effects.

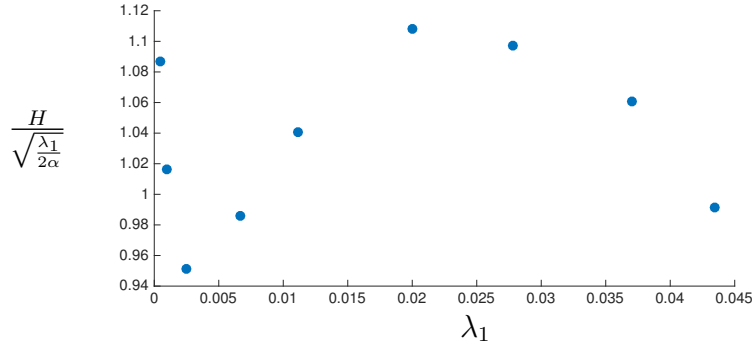


FIGURE 5. We compare the average mean projection distance H (defined as average of (3.2)) to $\sqrt{\frac{\lambda_1}{2\alpha}}$ (smoothing length scale) for the Experiment 3.3. We consider a somewhat broader set of λ_1 values than on Figure 4. We observe good agreement with the expectation, partly motivated by (3.4), that $H \sim \sqrt{\frac{\lambda_1}{2\alpha}}$.

the smallest scale at which distinct components can be detected by the (MPPC) functional. When distances are larger than λ_2 , connectedness is governed by the projected linear density α of the curves, as we investigate with the following simple example.

Example 3.5. Uniform density on line. In this example, we consider the measure μ to have uniform density α on the line segment $[0, L] \subset \mathbb{R}$. We relegate the technical details of the analysis to Appendix A; here we report the main conclusions. By (A.6) there is a critical density

$$\alpha^* = \left(\frac{4}{3}\right)^2 \frac{\lambda_1}{\lambda_2^2}$$

such that if $\alpha > \alpha^*$ then the minimizer γ has one component and is itself a line segment contained in $[0, L]$. It is straightforward to check that γ will be shorter than L by a length of $h = \sqrt{\lambda_1/\alpha}$ on each side. Note that at the endpoints $H^2 = h^2/3$, which is less than the upper bound at interior points predicted by (3.1).

On the other hand, if $\alpha < \alpha^*$ and L is long enough then the minimizer consists of regularly spaced points on $[0, L]$ with space between them approximately (because of finite size effects)

$$(3.5) \quad \text{gap} \approx 2 \left(\frac{3\lambda_1\lambda_2}{4\alpha} \right)^{\frac{1}{3}}.$$

An example of this scenario is provided in Figure 6.

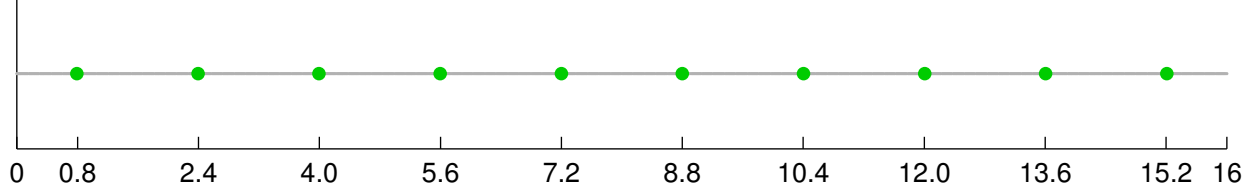


FIGURE 6. A minimizer for $n = 1000$ uniformly spaced points on a line segment, with total mass 1. Here $\lambda_1 = 1/16$, $\lambda_2 = .6$ and the critical value for connectedness is $\lambda_2^* = 4/3$. The optimal gap between the points is 1.6, compared to the approximation of ≈ 1.53 given by (3.5). The discrepancy is due to the finite length of the line segment considered in the example.

3.2. Summary of important quantities and length scales. Here we provide an overview of how length scales present in the minimizers are affected by the parameters λ_1 and λ_2 , and the geometric properties of data. We identify key quantities and length scales that govern the behavior of minimizers to (MPPC). We start with those that dictate the local geometry of penalized principal curves.

$\sqrt{\frac{\lambda_1}{2\alpha}}$ — smoothing length scale (discussed in Sections 3.1.1 and 3.1.2 and illustrated in Example 3.3). This scale represents the resolution at which data will be approximated by curves. Consider data generated by smooth curve with data density per length α and added noise (high-frequency oscillations, uniform distribution in a neighborhood of the curve, etc.). Then the noise will be "ignored" by the minimizer as long as its average amplitude (distance in space from the generating curve) is less than a constant multiple (depending on the type of noise) of $\sqrt{\frac{\lambda_1}{2\alpha}}$. In other words $\sqrt{\frac{\lambda_1}{2\alpha}}$ is the length scale over which the noise is averaged out. Noise below this scale is neglected by the minimizer, while noise above is interpreted as signal that needs to be approximated. For example, if we think of data as drawn by a pen, then $2\sqrt{\frac{\lambda_1}{2\alpha}}$ is the widest the pen tip can be, for the line drawn to be considered a line by the algorithm.

$\frac{\lambda_1\mathcal{K}}{\alpha}$ — bias or approximation-error length scale (discussed in Section 3.1.1). Consider again data generated by smooth curve with data density per length α and curvature \mathcal{K} . If the curvature of the curve is small (compared to $\sqrt{\frac{\lambda_1}{\alpha}}$) and reach is comparable to $1/\mathcal{K}$, then the distance from the curve to the minimizer is going to scale like $\frac{\lambda_1\mathcal{K}}{\alpha}$. That is the typical error in reconstruction of a smooth curve that a minimizer makes (due to the presence of the length penalty term) scales like $\frac{\lambda_1\mathcal{K}}{\alpha}$.

In addition to the above length scales, the following quantities govern the topology of multiple penalized principal curves:

λ_2 — connectivity threshold (discussed in Section 3.1.3). This length scale sets the minimum distance between distinct components of the solution. Gaps in the data of size λ_2 or less

are not detected by the minimizer. Furthermore, this quantity provides the scale over which the following critical density is recognized.

$\frac{\lambda_1}{\lambda_2}$ — linear density threshold (discussed in Example 3.5 and Appendix A.6). Consider again data generated (possibly with noise) by a smooth curve (with curvature small compared to $\sqrt{\frac{\lambda_1}{\alpha}}$) with data density per length α . If α is smaller than $\alpha^* = \left(\frac{4}{3}\right)^2 \frac{\lambda_1}{\lambda_2} + O(\mathcal{K})$, then it is cheaper for the data to be approximated by a series of points than by a continuous curve. That is if there are too few data points the functional no longer sees them as a continuous curve. If $\alpha > \alpha^*$, then the minimizers of (PPC) and (MPPC) are expected to coincide, while if $\alpha < \alpha^*$, then the minimizer of (MPPC) will consist of points spaced at distance about $\left(\frac{\lambda_1 \lambda_2}{\alpha}\right)^{\frac{1}{3}}$. Note that the condition $\alpha < \alpha^*$ can also be written as $\sqrt{\frac{\lambda_1}{\alpha}} < \frac{3}{4} \lambda_2$, and thus the minimizer can be expected to consist of more than component if the connectivity threshold is greater than the smoothing length scale.

We also remark the following scaling properties of the functionals. Note that $E_{a\mu}^{\lambda_1, \lambda_2} = a E_{\mu}^{\lambda_1/a, \lambda_2}$ for any $a > 0$. Thus, when the total mass of data points is changed, λ_1 should scale like $|\mu|$ to preserve minimizers. Alternatively, if $\mu_L(A) := \mu(\frac{A}{L})$ for every $A \subseteq \mathbb{R}^d$ and some $L > 0$, one easily obtains that $E_{\mu_L}^{\lambda_1, \lambda_2}(L\gamma) = L^2 E_{\mu}^{\lambda_1/L, \lambda_2/L}(\gamma)$.

3.3. Parameter selection. Understanding the length scales above can guide one in choosing the parameters λ_1, λ_2 . Here we discuss a couple of approaches to selecting these parameters. We will assume that the data measure μ has been normalized, so that it is a probability measure.

A natural quantity to specify is a critical density α^* , which ensures that the linear density of any found curve will be at least α^* . From Section 3.2 it follows that setting α^* imposes the following constraint on the parameters: $\frac{16}{9} \frac{\lambda_1}{\lambda_2^2} = \alpha^*$. Alternatively, one can set α^* if provided a bound on the desired curve length – if one is seeking a single curve with approximately constant linear density and length l or less, then set $\alpha^* = l^{-1}$.

There are a couple of ways of obtaining a second constraint, which in conjunction with the first determine values for λ_1, λ_2 .

3.3.1. Specifying critical density α^* and desired resolution H^* . One can set a desired resolution for minimizers by bounding the mean squared projection distance H . If α^* is set to equal the minimum of α along the curves then, the spatial resolution H from the data to minimizing curves is at most $\sqrt{\frac{\lambda_1}{2\alpha^*}}$. Consequently, if one specifies α^* and desires spatial resolution H^* , or better, the desired parameters are:

$$\lambda_1 = 2\alpha^* H^{*2} \quad \text{and} \quad \lambda_2 = \frac{4\sqrt{2}}{3} H^*.$$

Choosing proper H^* depends on the level of noise present in the data. In particular, H^* needs to be at least the mean squared height of vertical noise in order to prevent overfitting.

3.3.2. Specifying critical density α^* and λ_2 . One may be able to choose λ_2 directly, as it specifies the resolution for detecting distinct components. In particular, there needs to be a distance of at least λ_2 between components, in order for them to be detected as separate. Once set, $\lambda_1 = \frac{9}{16} \alpha^* \lambda_2^2$.

Typically one desires the smallest (best) resolution λ_2 , that does not lead to α^* larger than desired. Even if a single curve is sought, taking a smaller value for λ_2 can ensure less frequent undesirable local minima. One case of this is later illustrated in Example 4.1, where local minimizers can oscillate within the parabola.

Example 3.6. Line segments. Here we provide a simple illustration of the role of parameters, using data generated by three line segments with noise. The line segments are of the same length, and the ratio of the linear density of data over the segments is approximately 4:2:1 (left to right). In

addition, the first gap is larger than the second gap. Figure 7 shows how the minimizers of (MPPC) computed depend on parameters used. In the Subfigures 7(a), 7(b) 7(c) we keep λ_1 fixed while decreasing λ_2 . As the critical gap length is decreased, and equivalently having more components in the minimizer becomes cheaper, the gaps in the minimizer begin to appear. It no longer sees the data representing one line but two or three separate lines. the only difference between functionals in Subfigures 7(c) and Subfigures 7(d) is that λ_1 is increased from 0.008 to 0.024. This results in length of the curve becoming more expensive. In Subfigure 7(d) we see that, due to low data density per length (α), the minimizer approximates the two data patches to the right by singletons rather than curves.

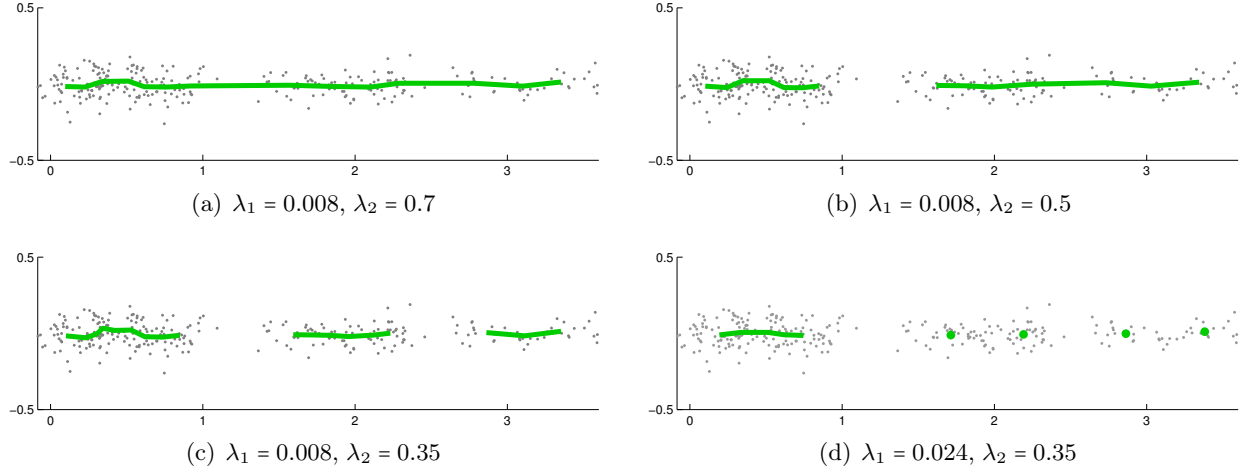


FIGURE 7. Minimizer of (MPPC) shown for different parameter settings. λ_1 and λ_2 .

4. NUMERICAL ALGORITHM FOR COMPUTING MULTIPLE PENALIZED PRINCIPAL CURVES

For this section we assume the data measure μ is discrete, with points $x_1, x_2, \dots, x_n \in \mathbb{R}^d$ and corresponding weights $w_1, w_2, \dots, w_n \geq 0$. The weights are uniform ($1/n$) for most applications, but we make note of our flexibility in this regard for cases when it is convenient to have otherwise.

For a piecewise linear curve $y = (y_1, \dots, y_m)$, we consider projections of data to y_i 's only. Hence, we approximate $d(x_i, y) \approx \min\{|x_i - y_j| : j = 1, \dots, m\}$, unless otherwise stated. (Notation: when a is a vector, as are x_i, y_j in the previous line, $|a|$ denotes the Euclidean norm). Before addressing minimization of (MPPC), we first consider (PPC) where y represents a single curve. The discrete form is

$$(4.1) \quad \sum_{j=1}^m \sum_{i \in I_j} w_i |x_i - y_j|^2 + \lambda_1 \sum_{j=1}^{m-1} |y_{j+1} - y_j|$$

where

$$(4.2) \quad I_j := \{i : (\forall k = 1, \dots, m) |x_i - y_j| \leq |x_i - y_k|\}$$

represents the set indexes of data points for which y_j is the closest among $\{y_1, \dots, y_m\}$. In case that the closest point is not unique an arbitrary assignment is made so that I_1, \dots, I_m partition $\{1, \dots, n\}$ (for example set $\tilde{I}_j = I_j \setminus \bigcup_{i=1}^{j-1} I_i$).

4.1. Basic approach for minimizing PPC. Here we restrict our attention to performing energy decreasing steps for the (PPC) functional. We emphasize again that this minimization problem is non-convex. The projection assignments I_1, \dots, I_m depend on y itself. However, if the projection assignments are fixed, then the resulting minimization problem is convex. This suggests the following EM-type algorithm outlined in Algorithm 1.

Algorithm 1 Computing local minimizer of (PPC)

Input: data x_1, \dots, x_n , weights w_1, \dots, w_n , initial curve y_1, \dots, y_m , $\lambda_1 > 0$
repeat
 1. compute I_1, \dots, I_m defined in (4.2)
 2. minimize (4.1) for I_1, \dots, I_m fixed as described in Section 4.1.1
until convergence

Note that if the minimization of (4.1) is solved exactly, then Algorithm 1 converges to a local minimum in finitely many steps (since there are finitely many projection states, which cannot be visited more than once).

4.1.1. Minimize functional with projections fixed. We now address the minimization of (4.1) with projections fixed (step 2 of Algorithm 1). One may observe that this subproblem resembles that of a regression, and in particular the fused lasso [38].

To perform the minimization we apply the alternating direction method of multipliers (ADMM) [3], which is equivalent to split Bregman iterations [22] when the constraints are linear (our case) [17]. We rewrite the total variation term as $\|Dy\|_{1,2} := \sum_{i=1}^{m-1} |(Dy)_i|$, where D is the difference operator, $(Dy)_i = y_{i+1} - y_i$ and $|\cdot|$ again denotes the Euclidean norm. An equivalent constrained minimization problem is then

$$\min_{y, z : z = Dy} \sum_{j=1}^m \sum_{i \in I_j} w_i |x_i - y_j|^2 + \lambda \|z\|_{1,2}$$

Expanding the quadratic term and neglecting the constant, we obtain

$$(4.3) \quad \min_{y, z : z = Dy} \|y\|_{\bar{w}}^2 - 2(y, \bar{x})_{\bar{w}} + \lambda \|z\|_{1,2}$$

where notation was introduced for total mass projecting to y_j by $\bar{w}_j = \sum_{i \in I_j} w_i$, center of mass $\bar{x}_j = \frac{1}{\bar{w}_j} \sum_{i \in I_j} w_i x_i$, and weighted inner product $(y, \bar{x})_{\bar{w}} = \sum_{j=1}^m \bar{w}_j (y_j, \bar{x}_j)$. One iteration of the ADMM algorithm then consists of the following updates:

- (1) $y^{k+1} = \operatorname{argmin}_y \|y\|_{\bar{w}}^2 - 2(y, \bar{x})_{\bar{w}} + \frac{\rho}{2} \|Dy - z^k + b^k\|^2$
- (2) $z^{k+1} = \operatorname{argmin}_z \lambda \|z\|_{1,2} + \frac{\rho}{2} \|Dy^{k+1} - z + b^k\|^2$
- (3) $b^{k+1} = b^k + Dy^{k+1} - z^{k+1}$

where $\rho > 0$ is a parameter that can be interpreted as penalizing violations of the constraint. As such, lower values of ρ tend to make the algorithm more adventurous, though the algorithm is known to converge to the optimum for any fixed value of $\rho > 0$.

The minimization in the first step is convex, and the first order conditions yield a tridiagonal system for y . The tridiagonal matrix to be inverted is the same for all subsequent iterations, so only one inversion is necessary, which can be done in $\mathcal{O}(md)$ time. In the second step, z decouples, and the resulting solution is given by block soft thresholding

$$z_i^{k+1} = \begin{cases} v_i^k - \frac{\lambda}{\rho} \frac{v_i^k}{\|v_i^k\|} & \text{if } \|v_i^k\| > \frac{\lambda}{\rho} \\ 0 & \text{else} \end{cases}$$

where we have let $v_i^k = (Dy^{k+1})_i + b_i^k$. We therefore see that ADMM applied to (4.3) is very fast.

Note that one only needs for the energy to decrease in this step for Algorithm 1 to converge to a local minimum. This is typically achieved after one iteration of ADMM. In such cases few iterations may be appropriate, as finer precision typically gets lost once projections are updated. On the other hand, the projection step is more expensive, requiring $\mathcal{O}(nmd)$ operations to compute exactly. It may be worthwhile to investigate how to optimize alternating these steps, as well as more efficient methods for updating projections especially when changes in y are small. In our implementation we exactly recompute all projections, and if the resulting change in energy is small, we minimize (4.1) to a higher degree of precision (apply more iterations of ADMM before again recomputing projections).

4.2. Approach to minimizing MPPC. We now discuss how we perform steps that decrease the energy of the modified functional (MPPC). We allow $y = y_1, \dots, y_m$ to consist of any number, k , of curves, and we denote them $y^1 = (y_1, \dots, y_{m_1})$, $y^2 = (y_{m_1+1}, \dots, y_{m_1+m_2})$, \dots , $y^k = (y_{m-m_k+1}, \dots, y_m)$, where $m_1 + m_2 + \dots + m_k = m$. The indexes of the curve ends are $s_c = \sum_{j=1}^c m_j$ for $c = 1, \dots, k$, and we set $s_0 = 0$. The discrete of form of (MPPC) can then be written as

$$(4.4) \quad \sum_{j=1}^m \sum_{i \in I_j} w_i |x_i - y_j|^2 + \lambda_1 \sum_{c=0}^{k-1} \sum_{j=1}^{m_{c+1}} |y_{s_c+j+1} - y_{s_c+j}| + \lambda_1 \lambda_2 (k-1).$$

Our approach to (locally) minimizing the problem over y, k, m_1, \dots, m_k is to split the functional into parts that are decreased over different variables. Keeping k, m_1, \dots, m_k constant and minimizing over y_1, \dots, y_m we can decrease (4.4) by simply applying step 1 and step 2 of Algorithm 1 to each curve y^i , $i = 1, \dots, k$ (note that step 2 can be run in parallel). To minimize over k, m_1, \dots, m_k we introduce topological routines below that disconnect, connect, add, and remove curves based on the resulting change in energy.

4.2.1. Disconnecting and connecting curves. Here we describe how to perform energy decreasing steps by connecting and disconnecting curves. We first examine the energy contribution of an edge $\{i, i'\} := [y_i, y_{i'}]$. To do so we compare the energies corresponding to whether or not the given edge exists. It is straightforward to check that the energy contribution of the edge $\{i, i'\}$ with respect to the continuum functional (MPPC) is

$$\Delta E_{i,i'} := \lambda_1 |y_{i'} - y_i| - \lambda_1 \lambda_2 - \sum_{j \in I_{i,i'}} w_j \min(|y_i - \Pi_{i,i'}(x_j)|, |y_{i'} - \Pi_{i,i'}(x_j)|)^2$$

where $I_{i,i'}$ is the set of data points projecting to the edge $\{i, i'\}$, and $\Pi_{i,i'}$ is the orthogonal projection onto edge $\{i, i'\}$. Our connecting and disconnecting routines will be based on the sign of $\Delta E_{i,i'}$. We note that above criterion is based on the variation of the continuum functional rather than its discretization (4.4), in which projections to the vertices only (not edges) are considered. Our slight deviation here is motivated by providing a stable criterion that is invariant to further discretizations of the line segment $[y_i, y_{i'}]$. While we use the discrete functional to simplify computations in approximating the optimal fitting of curves, we will connect and disconnect curves based on the continuum energy (MPPC).

We first discuss disconnecting. We compute the energy contribution for each existing edge and if $\Delta E_{i,i'} < 0$, then we remove edge $\{i, i'\}$. Note this condition can only be true if the length of the edge is at least λ_2 . It may happen that all edge lengths are less than λ_2 , but that the energy may be decreased by removing a sequence of edges, whose total length is greater than λ_2 . Thus, in addition to checking single edges, we implement an analogous check for sequences of edges. The energy contribution of a sequence of k edges $\{i, i+1\}, \{i+1, i+2\}, \dots, \{i+k-1, i+k\}$ (including the

corresponding interior vertices $y_{i+1}, \dots, y_{i+k-1}$) is given by

$$\begin{aligned} \Delta E_{i:i+k} := & \lambda_1 \left(\sum_{l=0}^{k-1} |y_{i+l+1} - y_{i+l}| - \lambda_2 \right) \\ & + \sum_{l=0}^{k-1} \sum_{j \in I_{i+l, i+l+1}} w_j \left((x_j - \Pi_{i+l, i+l+1}(x_j))^2 - (\min\{|x_j - y_i|, |x_j - y_{i+k}|\})^2 \right). \end{aligned}$$

The routine for checking such edge sequences is outlined in Algorithm 2.

Algorithm 2 Removing appropriate edge sequences

Input: data x_1, \dots, x_n , weights w_1, \dots, w_n , connected curve y_1, \dots, y_m , projections I , $\lambda_1, \lambda_2 > 0$
 set $i = 1$
repeat
 set $k = 1$, $len = |y_{i+1} - y_i|$
 repeat
 increment $k = k + 1$, $len = len + |y_{i+k} - y_{i+k-1}|$
 until $len > \lambda_2$ (or $i + k = m$, in which case break)
 compute $\Delta E_{i:i+k}$
 if $\Delta E_{i:i+k} < 0$ **then**
 remove edge sequence $\{i, i+1\}, \{i+1, i+2\}, \dots, \{i+k-1, i+k\}$
 advance $i = i + k - 1$
 end if
 increment $i = i + 1$
until $i > m - 1$

Connecting is again based on the energy contribution of potential new edges. We use a greedy approach to adding the edges. That is, we compute $\Delta E_{i,i'}$ for each potential edge $\{i, i'\}$, and add them in ascending order, connecting curves until no admissible energy-decreasing edges exist. We note that finding the globally optimal connections is essentially a traveling salesman problem, which is NP-hard. More sophisticated algorithms could be used here, but the greedy search is simple and has satisfactory performance.

4.2.2. Management of singletons: Here we describe the procedures for topological changes via adding and removing components of the multiple curves. This is achieved by adding singletons (curves whose range is just a single point in \mathbb{R}^d), growing them into curves, and by removing singletons. Even if one is only interested in recovering one-dimensional structures, singletons may play a vital role. In particular, any low-density regions of the data (background noise or outliers) can often be represented by singletons in a minimizer of (MPPC), allowing the curves to be much less affected in approximating the underlying one-dimensional structure.

Below we provide effective routines for energy-decreasing transitions between configurations involving singletons. For checking whether (and where) singletons should be added, we examine each point y_i individually. If y_i is itself not a singleton, we compute the expected change in energy resulting from disconnecting y_i from its curve, placing it at the mean \bar{x}_i of the data that project to it, and reconnecting the neighbors of y_i , so the number of components only increases by one. The change in the fidelity term will be exactly $-\bar{w}_i(\bar{x}_i - y_i)^2$, where $\bar{w}_i = \sum_{j \in I_j} w_j$ is the total mass projecting to y_i . Thus we add a singleton when

$$\lambda_1 \lambda_2 < \bar{w}_i(\bar{x}_i - y_i)^2 + \lambda_1 (|y_i - y_{\max(1, i-1)}| + |y_i - y_{\min(m, i+1)}| - |y_{\max(1, i-1)} - y_{\min(m, i+1)}|).$$

If y_i is itself a singleton, then one cannot exactly compute the change in the energy due to adding another singleton in its neighborhood without knowing the optimal positions of both singletons. We

restrict our attention to the data which project onto y_i , and note that if those points are the only ones that project to the new singleton, then adding the singleton may be advantageous only if the fidelity term associated to y_i is greater than $\lambda_1 \lambda_2$. If that holds, we perturb y_i in the direction of one of its data points, place a new singleton opposite to y_i with respect to its original position, and apply a few iterations of Lloyd's k-means algorithm (with $k = 2$) to the data points that projected to y_i . We keep the two new points if and only if the energy decreases below that of the starting configuration with only y_i .

A singleton y_i gets removed if doing so decreases the energy. That is if

$$\lambda_1 \lambda_2 > \sum_{j \in I_i} w_j (|x_j - y_i|^2 - d(x_j, y_{-i})^2)$$

where $d(x_j, y_{-i}) := \min\{|x_j - y_{i'}| : i' \in [m], i' \neq i\}$.

Since singletons are represented by just a single point and cannot grow by themselves, we also check whether transitioning from singleton to short curve is advantageous. To do so we enforce that the average projection distance \tilde{d}_i to a singleton y_i is less than $\sqrt{\lambda_1/\tilde{\alpha}}$, which represents the expected spatial resolution, where $\tilde{\alpha} = \bar{w}_i/(4\tilde{d}_i)$ is an approximation to the potential linear density. Thus we add a neighboring point to y_i if

$$\tilde{d}_i = \sum_{j \in I_i} w_j |x_j - y_i| > \lambda_1/\bar{w}_i.$$

Since this is based on an approximation, we also explicitly compute the posterior energy to make sure that it has indeed decreased, and only in this case keep the change.

Note that for each singleton y_j , minimizing the discrete energy (4.1) with projections fixed corresponds to placing y_j at its center of projected mass \bar{w}_i . Hence for singletons Algorithm 1 reduces to Lloyd's k-means algorithm.

In summary, we have fast and simple ways to perform energy decreasing steps involving the λ_2 term of the functional. Even when minimizers are expected to be connected, performing these steps may change the topological structure of the curve, keeping it in higher density regions of the data, and consequently evading several potential local minima of the original functional (PPC).

4.3. Re-parametrization of y . In applying the algorithm described thus far, it may, and often does, occur that some regions of y are represented with fewer points y_i than others, even if an equal amount of data are projected to those regions. That is, there is nothing that forces the nodes y_i to be well spaced along the discretized curve. To address this, we introduce criteria that $l_i \bar{w}_i$ be roughly constant for $i = 1, \dots, m$, where $l_i = \frac{1}{2} \sum \{|y_i - y_j| : j \in \{i-1, i+1\} \cap [1, m]\}$ and \bar{w}_i is the total weight of points projecting to y_i . This condition is motivated by finding for fixed m the optimal spacing of y_i 's that minimizes the fidelity term of the discrete energy (4.4), under the assumption that the data are distributed with slowly changing density in a rectangular tube around straight line y .

4.4. Criteria for well-resolved curves. Here we discuss criteria for when a curve can be considered well-resolved with regard to the number of points m used to represent it. One would like to have an idea of what conditions give an acceptable degree of resolution, without requiring m too large and significantly increasing computational time. We suggest two such conditions.

One is related to the objective of obtaining an accurate topological representation of the minimizer, specifically the number of components. In order to have confidence in recovering components at a scale λ_2 , the spacing between consecutive points on a discretized curve should be of the same scale. Thus we impose that the average of the edge lengths is at most $\frac{\lambda_2}{2}$.

Another approach for determining the degree of resolution of a curve is to consider its curvature. One may calculate the average turning angle and desire that it be less than some value (e.g. $\frac{\pi}{10}$). If λ_2 is not small enough, the first condition will not guarantee small turning angles, and so we include this criterion as optional in our implementation. We note that in light of the possible lack

of regularity of minimizers [35], it would not be reasonable to limit the maximal possible turning angle.

If either of the above criteria are not satisfied, we add more points to the curves where we expect they would decrease the discrete energy the most. Consistent with the criteria above for re-parametrization of the curves, we add points along the curve where $l_i \bar{w}_i$ is the largest.

4.5. Initialization. Finally, we discuss initialization. While the procedures described above enable the algorithm to evade many undesirable local minima, initialization can still impact the quality of the computed local minimizers. One of the simple ideas that we found to work very well is to initialize using singletons. We note that when the number of singletons is a fixed number k then minimizing (MPPC) reduces to minimizing the k -means functional. Thus to position the singletons for fixed k we use the standard Lloyd's algorithm to find the k -means cluster centers. We denote the (MPPC) energy of the k -means centers by $E(k)$. To determine a suitable value of k we perform a line search by starting with $k = 1$ and double it as long as $E(k)$ decreases, and then halve the intervals until a (local) minimizer k is found. We list the steps in Algorithm 3.

Algorithm 3 Initializing with singletons

Input: data x_1, \dots, x_n , weights w_1, \dots, w_n , and $\lambda_1, \lambda_2 > 0$.
Set $k = 1/2$, $E(k) = +\infty$
repeat
 Let $k = 2k$
 Compute the k -means centers $C_k = \{c_1, \dots, c_k\}$, and energy $E(k) := E_{\mu}^{\lambda_1, \lambda_2}(C_k)$
until $E(k) > E(k/2)$
Let $k' = \lfloor \frac{k}{2} \rfloor$, $k'' = k'$
repeat
 Let $k'' = \lfloor \frac{k''}{2} \rfloor$, $k = k' + k''$
 Compute the k -means centers $C_k = \{c_1, \dots, c_k\}$, and energy $E(k) := E_{\mu}^{\lambda_1, \lambda_2}(C_k)$
 if $E(k) < E(k')$ **then**
 $k' = k$
 end if
until $k'' = 1$
Output: $y = C_{k'}$

4.6. Overview. Thus far we have described all of the main pieces of our algorithm to compute local minimizers of (MPPC). Here we describe how we put these pieces together. Algorithm 1, which includes ADMM for decreasing the discrete energy (4.1), computes approximate local minimizers of (PPC). To approximate local minimizers of (MPPC), we break up the minimization into separate parts. One consists of a “local” step that updates the placement of each curve, and is accomplished by running the ADMM step of Algorithm 1 on each curve. On the other hand, the inclusion of routines to disconnect, connect, add, and remove curves allows us to perform energy-decreasing steps of (MPPC) in a more global topological fashion.

We provide an general outline for finding local minimizers of (MPPC) in Algorithm 4. The (potentially) topology-changing routines outlined in 4.2.1, 4.2.2 are run on a regular basis throughout the steps of Algorithm 1. In particular we run them every $top_period = 10$ iterations and we run the reparameterization of curves every $reparam_period = 5$ iterations. The performance for different values, as well as for different order of operations was similar.

Algorithm 4 Computing local minimizer of (MPPC) [Main Loop]

Input: data x_1, \dots, x_n , weights w_1, \dots, w_n , initial curve y_1, \dots, y_m , $\lambda_1, \lambda_2 > 0$.
 set iter = 0
repeat
 1. iter = iter + 1
 2. compute I_1, \dots, I_m , defined in (4.2)
 3. run ADMM on non-singleton curves to decrease energy (4.4) as described in Section 4.1.1
 4. replace y_j by the center of mass of data points projecting to y_j : $y_j = \bar{x}_j$
 if iter + 4 = 0 (mod top_period) **then**
 remove appropriate edge sequences as described in Section 4.2.1 and Algorithm 2
 else if iter + 3 = 0 (mod top_period) **then**
 add or remove appropriate singletons as described in Section 4.2.2
 else if iter + 2 = 0 (mod top_period) **then**
 add appropriate connections as described in Section 4.2.1
 else if iter + 1 = 0 (mod reparam_period) **then**
 add points and re-parametrize the curves if needed as described in Sections 4.3, 4.4
 end if
until convergence

4.7. Further numerical examples. We present a couple of further computational examples which illustrate the behavior of the functionals and the algorithm. For some of the examples, we include comparisons with results from other approaches including the Subspace Constrained Mean Shift algorithm and diffusion maps.

Example 4.1. Parabola. We begin with an example that illustrates the cutting and reconnecting mechanism used in the Algorithm 4 for finding minimizers of (MPPC). We use data that are uniformly distributed on the graph of the parabola $x = y^2$ for $y \in [-3, 3]$ and set $\lambda_1 = 0.12$ and $\lambda_2 = 4/3$. For illustration, we first run the Algorithm 4 for minimizing (PPC) (the same as main loop of Algorithm 4 without allowing any topological changes) starting from a small perturbation of the line segment $[0, 9] \times \{0\}$. The result is shown on Figure 8(a). We then turned on the cutting routine, described in Algorithm 2. The segments to be cut are indicated on Figure 8(a) as dashed lines. Figure 8(b) shows a subsequent configuration, after a few steps of ADMM relaxation, but prior to reconnecting. Edges that are about to be added in the reconnection step (described in Section 4.2.1) are shown as dashed blue lines.

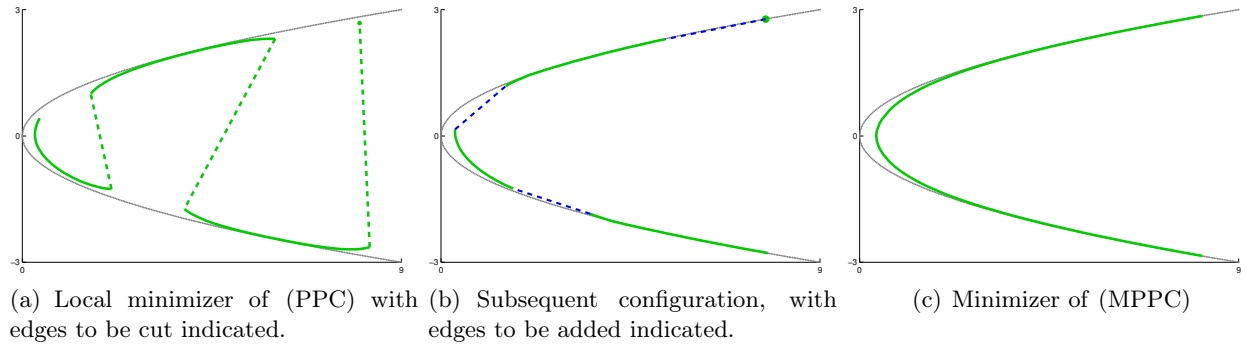


FIGURE 8.

Example 4.2. Noisy spiral. Here we consider data generated as noisy samples of the spiral $t \mapsto (t \cos(t), t \sin(t))$, $t \in [3, 14]$, shown as a dashed line in Figure 9(b). 2000 points are drawn uniformly with respect to arc length along the spiral. For each of these points, noise drawn independently from the normal distribution $1.5\mathcal{N}_2(0, 1)$ is added. In Figure 9, we show the results of algorithms for minimizing (PPC) and (MPPC). The initialization used for both experiments is a diagonal line corresponding to the first principal component. The descent for (PPC) does not allow for topological changes of the curve and subsequently gets attracted to a local minimum. Meanwhile, Algorithm 4 for minimizing (MPPC) is able to recover the geometry of the data, via disconnecting and reconnecting the initial curve.

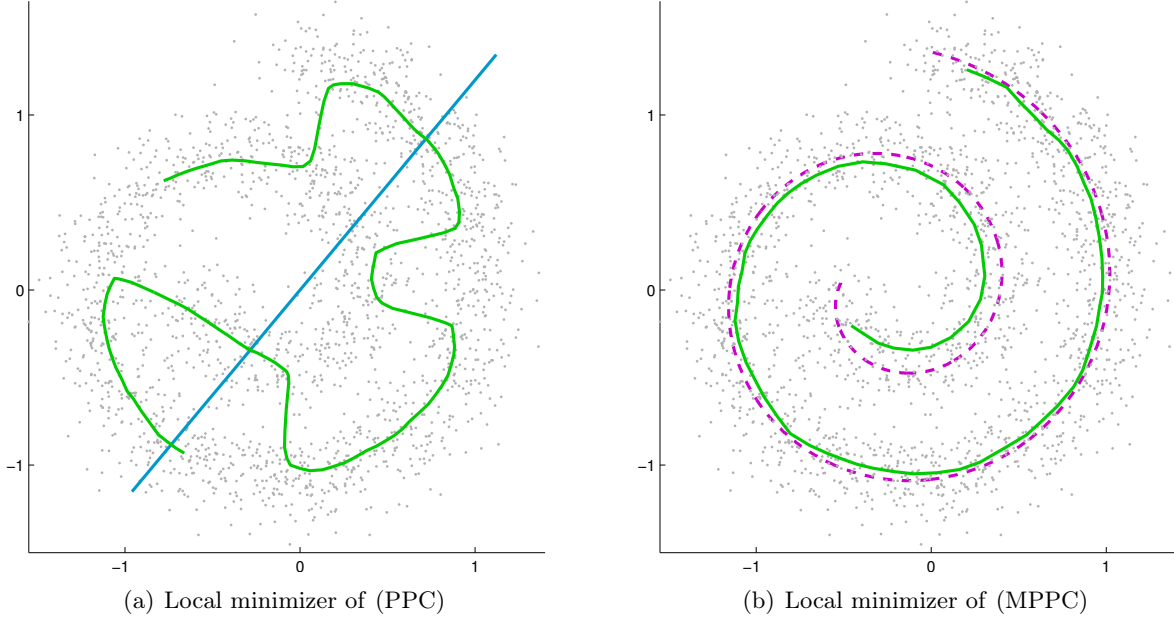


FIGURE 9. Numerical results on data generated by a spiral (in purple) plus Gaussian noise. On the left, (PPC) is used to find the local minimizer given by the green curve, using the first principal component (blue) as initialization. On the right, (MPPC) (with the same initialization) is minimized to find the green curve, using critical linear density $\alpha^* = .09$. In both cases $\lambda_1 = .01$.

For this dataset we also include results of the Subspace Constrained Mean Shift (SCMS) algorithm [31], also studied in [8, 9, 20] as means to find one dimensional structure in data. SCMS seeks to find the ridges (of an estimate) of the underlying probability density of the data. The ridge set of a function $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is defined as the set where ∇F is an eigenvector of the Hessian of F and the eigenvalues of all remaining eigenvectors are negative (the point is a local maximum along all orthogonal directions). In practice, given a random sample one uses a kernel density estimator (KDE) to approximate the probability distribution. SCMS algorithm takes a set of points as input, and successively updates each point until it converges to a ridge point of the KDE of a specified bandwidth. The output is then a list of unordered points that approximate the ridge set. We apply SCMS using a Gaussian kernel density estimate (KDE) for two different bandwidth, which both give good results. The algorithm is initialized with 2500 points on a 50×50 mesh in the range of the data. As shown in Figure 10, the algorithm does output points approximating the underlying one-dimensional structure. We note however that mathematically there is a number of ridges of

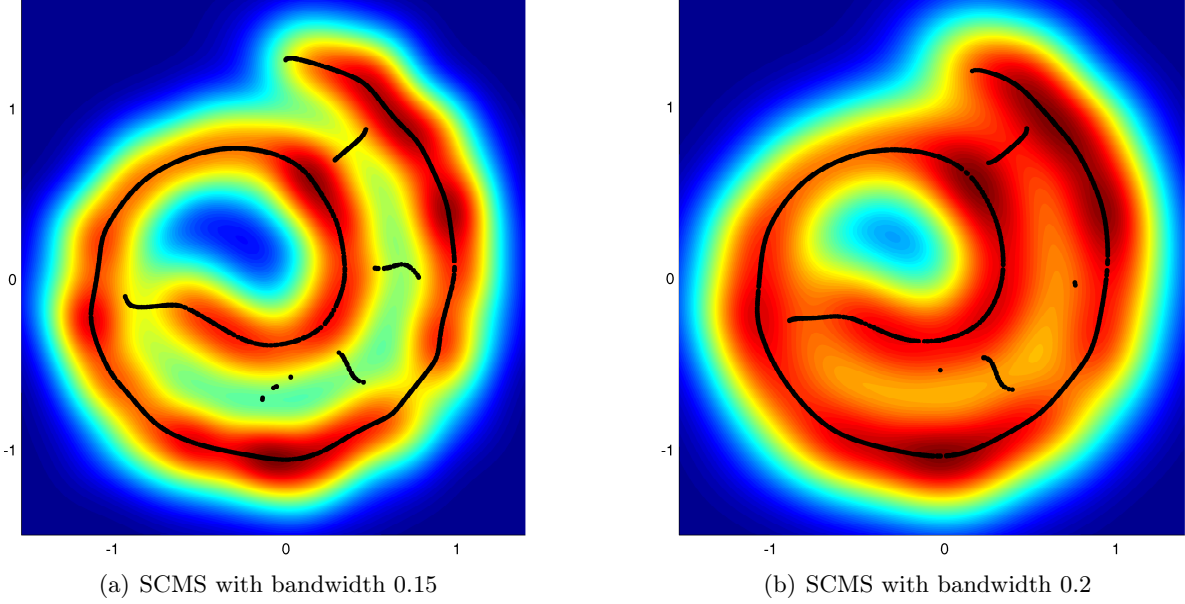


FIGURE 10. There are more undesirable ridges for small bandwidths. Decreasing the bandwidth further can also result in the gaps in the desirable spiral filament. There are fewer undesirable ridges at higher bandwidth, but they have higher density. Increasing the bandwidth further introduces a significant bias of the main ridge, compared to the generating curve.

KDE going between the layers of the spiral. The SCMC algorithm captures those with high enough density and large enough "basin of attraction". We note that as the kernel bandwidth increases, the number of undesirable ridges decreases, but their intensity increases (the density at the remaining ridges is higher). Removing points on the mesh that have density below a given threshold has been suggested for noisy data [9, 20], and doing so can improve the results here by eliminating some of the undesirable ridges. However, this introduces a parameter (density threshold) that needs to be chosen carefully (see appendix A of [9]).

Example 4.3. Noisy grid with background clutter. In the following example we illustrate the robustness of the proposed approach to background noise. We use data in \mathbb{R}^3 with an underlying grid-like structure, shown in Figure 11. The data consist of 2400 points generated by four intersecting lines with Gaussian noise, plus 2400 more points uniformly sampled from the background $[0, 3] \times [0, 3] \times [-.75, .75]$.

Since the linear density of data in the background noise is less than that of the intersecting lines, the computed minimizer approximates the data in the background by isolated points (in green). For the parameters we used, this is predicted by the discussion of the density threshold in Section 3. By choosing λ_1 and λ_2 so that the critical density threshold $\alpha^* = \left(\frac{4}{3}\right)^2 \frac{\lambda_1}{\lambda_2^2}$ is between the linear density of the background noise and the linear density of the lines, the background noise will be represented by isolated points, which allows the curves to appropriately approximate the intersecting lines.

We note that although the algorithm succeeds in approximating the one-dimensional structure of the data, it is not able to recover the intersections due to the simpler structure of configurations we consider in (MPPC). In such cases where our approach cannot identify the global topology, we presume it may be possible to use the obtained approximation as input for other approaches that aim to recover the topology of the data [34].

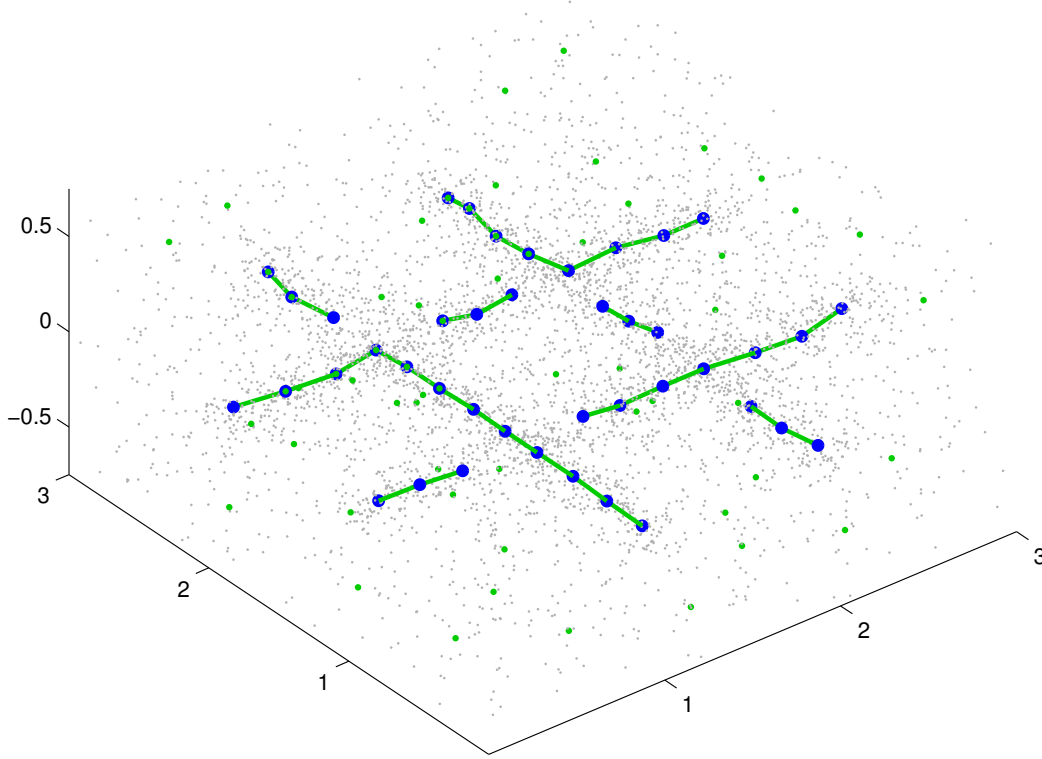


FIGURE 11. Computed minimizer on data generated from four intersecting lines forming a grid with Gaussian noise and background clutter in \mathbb{R}^3 , with $\lambda_1 = 7 \times 10^{-4}$, $\lambda_2 = 0.2$. The minimizer consists of curves and isolated points (green). The larger blue dots represent the discretization (points y_i) of curves which are a part of the minimizer.

Example 4.4. Zebrafish embryo images. Here we demonstrate performance of the algorithm on a high-dimensional dataset that consists of grayscale images.

In [14] Dsilva et al. develop a technique for finding the temporal order of still images of a developmental process. They consider the problem where both the time ordering and the angular orientation of the images are unknown. To be able to handle both variables simultaneously they use vector diffusion maps [33]. One of the tests they performed to validate their approach was on images taken from a time-lapse movie that captures zebrafish embryogenesis [https://zfin.org/zf_info/movies/Zebrafish.mov] (Karlstrom and Kane [24]).

In this case the angle of rotation is fixed, recovering the temporal order can be done using diffusion maps [10] alone, see Figure 13(b). Here we demonstrate that these images can also be ordered using our method.

As in [14], we apply our algorithm to 120 consecutive frames (roughly corresponding to seconds 6-17 in the movie) of 100x100 pixels in order to test how well it can recover the development trajectory. Thus each image is represented as a point in $\mathbb{R}^{10,000}$. We note that there is almost no noise in the dataset, but emphasize that the goal here is to recover a single curve passing through data whose true order is not provided to the algorithm.

After normalizing the data, we run our algorithm with parameters $\lambda_1 = 10^{-3}$ and $\lambda_2 = 2$. The low value for λ_1 is appropriate given that there is virtually no noise. The high value of λ_2 ensures that a single curve is found, and so the functional (PPC) is also being minimized. Our algorithm outputs a curve that correctly ranks all of the original images. Figure 12 shows a random sample

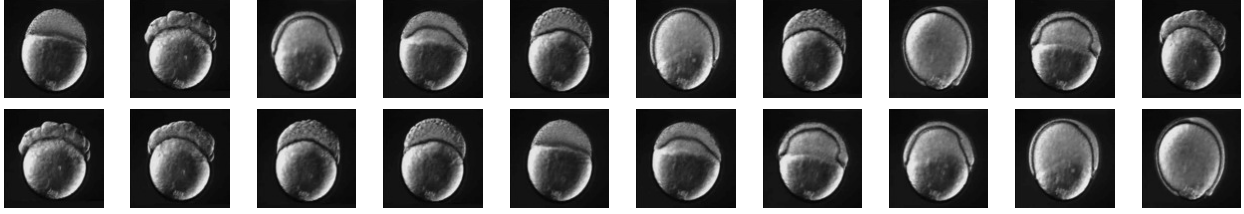
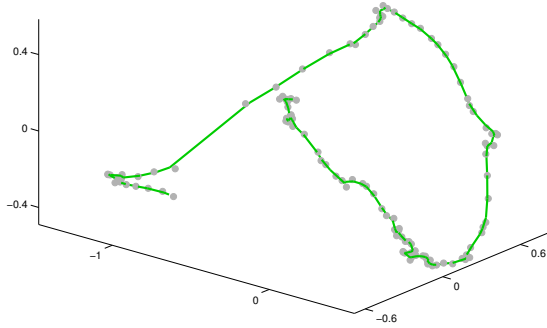
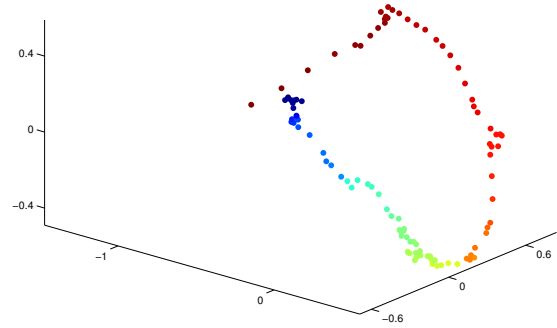


FIGURE 12. The top row shows 10 images of zebrafish embryos in random order. The bottom row shows 10 images ordered by the found curve that minimizes (MPPC).

of the images used, along with their found true ordering. In Figure 13(a), we visualize the found curve in \mathbb{R}^3 using the first three principal components.



(a) The curve found minimizing (MPPC)



(b) Color-coded first embedding coordinate of the diffusion map

FIGURE 13. On both images the first three principal components are used for visualization. The (MPPC) algorithm was applied to all 120 images, while we applied diffusion maps to only the first 104 images due to a slight camera shift that resulted in relatively large euclidean distance between images 104 and 105. Both methods perfectly ranked their respective data, and some (simple) preprocessing done in [14] allows diffusion maps to work on the full 120 images.

Example 4.5. Noisy spiral revisited. In the previous example we discussed the feasibility of using nonlinear dimensionality reduction techniques such as diffusion maps to order the data. Since the data in Example 4.4 had almost no noise, one can obtain a good ordering using many different methods. Spectral dimensionality reduction techniques are often successful even when substantial noise is present. However, when there is significant overlap in the distribution of data whose generating points have large intrinsic distance, spectral methods can fail to recover the desired one-dimensional ordering. The example below illustrates this and indicates that in some situations minimizing MPPC gives better results in ordering the data than diffusion maps.

We revisit the noisy spiral data considered in Example 4.2, and run the diffusion maps algorithm using a range of scaling parameters $\epsilon = (\frac{d}{c})^2$, where d denotes the median of the pairwise distances of the data points. After testing a wide range of parameter values c , we found that for all values tested the spectral embedding fails to recover the desired one-dimensional ordering. We display the

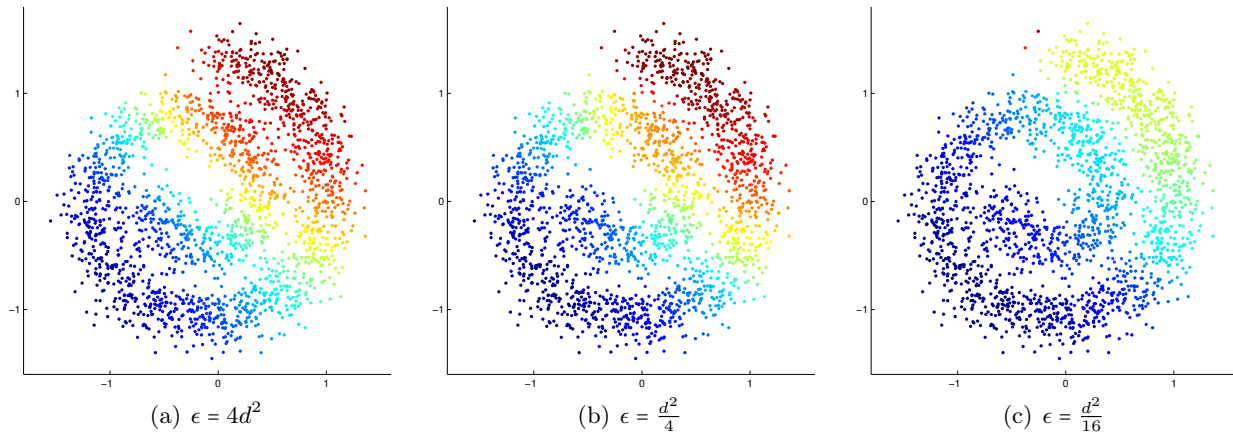


FIGURE 14. Color-coded one-dimensional embeddings provided by the diffusion maps algorithm for three settings of the scaling parameter ϵ . No values for ϵ recover the intrinsic ordering of the data. Larger values (left and middle) cannot detect the finer structure, while a smaller value (right) separates two outliers at the top (colored red and brown) from the rest of the data.

typical results (which correspond to $c = 0.5, 2$, and 4) in Figure 14. Larger values of ϵ lead to an embedding that differentiates the data linearly from bottom left to top right, while smaller values lead to an embedding that separates outliers in the top from the rest of the data points. On the other hand minimizer of (MPPC) can correctly recover the one dimensional structure, as shown in Figure 9(b).

5. DISCUSSION AND CONCLUSIONS

In this paper, we proposed a new objective functional (MPPC) for finding one-dimensional structures in data that allows for representation consisting of several components. The functional introduced is based on the average-distance functional and can be seen as a regularization of principal curves. It penalizes the approximation error, total length of the curves, and the number of curves used. We have investigated the relationship between the data generated by one dimensional signal with noise, the parameters of the functional, and the minimizer. Our findings provide guidance for the choice of parameters, and can further be used for multi-scale representation of the data. In addition, we have demonstrated that the zeroth-order term helps energy descent based algorithms converge to desirable configurations. In particular, energy descent approaches for (PPC) very often end up in undesirable local minima. The main reason for this is of topological nature – points on the approximate local minimizer represent the data points in an order which may be very different from the true ordering corresponding to the (unknown) generating curve. The added flexibility of being able to split and reconnect the curves provides a way for resolving such topological obstacles.

Finally, we have developed a fast numerical algorithm for estimating minimizers of (MPPC). It has computational complexity $\mathcal{O}(mnd)$, where n is the number of data points in \mathbb{R}^d , and m is the number of points along the approximating curve(s). We demonstrated the effectiveness of the algorithm in recovering the underlying one-dimensional structure for real and synthetic data, in cases with significant noise and in very high dimensions.

5.1. Relation to other approaches. We now briefly compare the proposed approach to other existing approaches for finding one-dimensional structures in data. The original principal curves are prone to overfitting, as carefully explained in [21] and are difficult to compute numerically.

The approach of Gerber and Whitaker [21] offers a more stable notion and an effective numerical implementation in low dimensions. Experiments by the authors indicate that the algorithm often selects desirable minimizers. However, the functional may still overfit noisy data, as there are many curves which minimize the functional but do not represent the data well. We also note that since the functional does not measure the approximation error, its low values are not a measure of how well the data are approximated. On the other hand if the data are sampled from a smooth curve the minimizers can be expected to recover the curve exactly.

A number of works inspired by principal curves treat the problem by considering objective functionals which regularize the principal curves problem. Among them are works of Tibshirani [37] (square curvature penalization) Kegl, Krzyzak, Linder, and Zeger [25] (length constraint), Biau and Fischer [2] (length constraint) Smola, Mika, Schölkopf, and Williamson [36] (a variety of penalizations including penalizing length as in (PPC)). These works are similar in spirit to our approach. The work of Biau and Fischer [2] also discusses model-selection based automated ways to choose parameters of the given functional for the specific data set. Wang and Lee [40] also use model selection to select parameters, but ensure the regularity of the minimizer in a different way. Namely they model the points along the curve as an autoregressive series.

Regarding numerical approaches, Kegl, Krzyzak, Linder, and Zeger [25] proposed a polygonal-line algorithm that penalizes sharp angles. Feuersänger and Griebel employ sparse grids to minimize a functional similar to (PPC), with length squared regularization [18] (as in [36]) for manifolds up to dimension three. While these approaches take measures against overfitting data, they do not address the problem of local minima, resulting in performance that is very sensitive to the initialization of the algorithms. Verbeek, Vlassis and Kröse [39] approach this issue by iteratively inserting, fitting, and connecting line segments in the data. This approach is effective in some situations where others exhibit poor performance (e.g. spiral in 2-d, some self-intersecting curves, and curvy data with little noise). However, in cases of higher noise the algorithm overfits if the number of segments is not significantly limited. A better understanding of the impact of the number of segments on the final configuration is still needed, despite some efforts to automatize selection of this parameter [40].

A different class of approaches to finding one dimensional structures is based on estimating the probability density function of the point cloud and then finding its ridges [16]. In particular Ozertem and Erdogmus [31] introduced the widely used Subspace Constrained Mean Shift (SCMS) algorithm, which is based on the Mean Shift algorithm of Comaniciu and Meer [11]. Estimation of density ridges has been substantially developed and studied – see works of Chen, Genovese, and Wasserman [8], Genovese, Perone-Pacífico, Verdinelli, and Wasserman [20], and Pulkkinen [32]. The SCMS algorithm is often able to find the desired one-dimensional structure even when significant noise is present. However it does not automatically parameterize the found one dimensional structure, which consists of an (unordered) set of points. An important difference between our approach and SCMC is that we seek the one-dimensional structure that best approximates the data, and measure the quality of approximation as part of the (MPPC) functional, while SCMC does not require the ridges found to approximate that data well. Let us also mention that in high dimensions SCMC faces a combination of computational difficulties, the primary of which is accurately estimating the Hessian of the density function (found using a kernel density estimator), as is discussed in Section 3 of [31].

Recently there has been a significant effort to recover one-dimensional structures which are branching and intersecting and in particular the connectivity network of the data set. The ability to recover graph structures and the topology of the data is very valuable, and facilitates a number of data analysis tasks. Several notable works are based on Reeb graphs and related structures [7, 19, 34]. We note that these approaches are sensitive to noise and furthermore the presence of noise significantly slows down the algorithms. We believe our approach and algorithm could be valuable as a pre-processing step for simplifying the data prior to applying graph-based approaches that find the connectivity network of the data set. Recalling the data from Example 4.3 and Figure

4.7, we see that although our approach does not recover the topological structure, it does identify and appropriately simplifies the one dimensional structure present in the data. The approaches mentioned here should work much better on the simplified (green or blue) data than on the original point cloud.

Finally we mention the work of Arias-Castro, Donoho, and Huo [1] who studied optimal conditions and algorithms for detecting (sufficiently smooth) one-dimensional structures with uniform noise in the background.

ACKNOWLEDGEMENTS

We are grateful to Xin Yang Lu, Ryan Tibshirani, and Larry Wasserman for valuable discussions. The research for this work has been supported by the National Science Foundation under grants CIF 1421502 and DMS 1516677. We are furthermore thankful to Center for Nonlinear Analysis (CNA) for its support.

APPENDIX A. ANALYSIS OF THE UNIFORMLY DISTRIBUTED DATA ON A LINE SEGMENT

Consider data uniformly distributed with density α on a line segment $[0, L]$. The functional (MPPC) takes the form

$$(A.1) \quad E_{\mu}^{\lambda_1, \lambda_2}(\gamma) := \int_0^L d(x, \gamma)^p \alpha dx + \lambda_1(L(\gamma) + \lambda_2 k(\gamma))$$

where $k(\gamma)$ is the number of components of γ minus 1. We restrict ourselves to γ such that $\{0, L\} \subset \text{range}(\gamma)$, so that γ takes the form $\gamma = \bigcup_{i=1}^{k+1} [a_i, b_i]$, where $a_1 = 0$, and $b_{k+1} = L$. Define $\tau := \sum_{i=1}^{k+1} \tau_i$, $g := \sum_{i=1}^k g_i$, where $\tau_i := b_i - a_i$ and $g_i := a_{i+1} - b_i$. We make the following observations:

Lemma A.1. *The energy $E_{\mu}^{\lambda_1, \lambda_2}$ is invariant under redistribution of total length of γ , assuming that the number of components is $k+1$, and that the gap sizes remain constant. More precisely, if $\bar{\gamma} = \bigcup_{i=1}^{k+1} [\bar{a}_i, \bar{b}_i]$, $\tilde{\gamma} = \bigcup_{i=1}^{k+1} [\tilde{a}_i, \tilde{b}_i]$ and there exists a permutation σ of $\{1, \dots, k\}$ such that $\bar{g}_i = \tilde{g}_{\sigma(i)}$ for $i = 1, \dots, k$, then $E_{\mu}^{\lambda_1, \lambda_2}(\bar{\gamma}) = E_{\mu}^{\lambda_1, \lambda_2}(\tilde{\gamma})$.*

Lemma A.2. *For $k > 0$ fixed, the energy $E_{\mu}^{\lambda_1, \lambda_2}$ is minimized when the length of the gaps between components are uniform. More precisely, consider an arbitrary $\gamma = \bigcup_{i=1}^{k+1} [a_i, b_i]$, with total gap g defined as above. Let $\tilde{\gamma}$ have $k+1$ components such that $\tilde{g}_i = g/k$, implying that $\tilde{g} = g$. Then $E_{\mu}^{\lambda_1, \lambda_2}(\tilde{\gamma}) \leq E_{\mu}^{\lambda_1, \lambda_2}(\gamma)$, with equality only if $g_i = \tilde{g}_i$.*

Proof. The result is trivial for $k = 0$. We prove the result for $k = 1$. Consider γ with $g = g_1 + g_2$. The fidelity part of the energy $E_{\mu}^{\lambda_1, \lambda_2}$, as a function of g_1 is

$$F(g_1) = 2 \int_0^{g_1/2} x^p \alpha dx + 2 \int_0^{(g-g_1)/2} x^p \alpha dx.$$

Thus

$$\frac{dF}{dg_1} = \frac{1}{2^{p-1}} \alpha (g_1^p - (g - g_1)^p)$$

and

$$\frac{d^2 F}{dg_1^2} = \frac{p}{2^{p-1}} \alpha (g_1^{p-1} + (g - g_1)^{p-1}) \geq 0.$$

By these we see that g_1 minimizes the energy if and only if $g_1 = g/2 = g_2$. The result for $k > 2$ follows since one can consider the above situation by looking at the gaps formed by three consecutive components. \square

Using Lemma A.1 we may assume that each component not containing the endpoints 0 or L has the same length l , and that the two components containing the endpoints are of length $l/2$. By Lemma A.2, the gaps between the components are $\frac{L-kl}{k}$. We first consider $k > 0$ fixed, and minimize the energy w.r.t. l in the range $l \in [0, \frac{L}{k})$. The energy

$$(A.2) \quad \begin{aligned} E &= \lambda_1 k l + 2k \int_0^{\frac{L-kl}{2k}} x^p \alpha dx + \lambda_1 \lambda_2 k \\ &= \lambda_1 k l + \frac{2}{p+1} k \left(\frac{L-kl}{2k} \right)^{p+1} \alpha + \lambda_1 \lambda_2 k \end{aligned}$$

is convex on $[0, \frac{L}{k})$. Taking a derivative in l we obtain

$$\frac{dE}{dl} = \lambda_1 k - k \left(\frac{L-kl}{2k} \right)^p \alpha.$$

Setting the derivative to zero and solving for l , and by noting that if there is no solution on $[0, \frac{L}{k})$ then E is a nondecreasing function of l , we get that the energy is minimized at

$$(A.3) \quad l_k^* = \begin{cases} \frac{L}{k} - 2 \left(\frac{\lambda_1}{\alpha} \right)^{1/p} & \text{if } k \leq \frac{L}{2 \left(\frac{\lambda_1}{\alpha} \right)^{1/p}} =: \bar{k} \\ 0 & \text{else.} \end{cases}$$

as we indicate above let $\bar{k} = 1 + \frac{L}{2 \left(\frac{\lambda_1}{\alpha} \right)^{1/p}}$. For k between 1 and \bar{k} , plugging back into (A.2) we get that the minimal energy is

$$(A.4) \quad E_{min}(k) = \lambda_1 L + \lambda_1 \lambda_2 k - \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p} k$$

By direct inspection we verify that (A.4) is the (minimal) energy in the case that there is only one component (no breaks in the line). We note that (A.4) is linear in k , and hence for k between 0 and \bar{k} , the minimizing value is at a boundary:

$$k^* = \begin{cases} 0 & \text{if } \lambda_2 \geq \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p} \\ \lfloor \bar{k} \rfloor & \text{otherwise} \end{cases}$$

We now consider $k > \bar{k}$ when all components have length zero ($l_k^* = 0$). The energy in this case is

$$E_{l=0}(k) := \frac{2}{p+1} \left(\frac{L}{2} \right)^{p+1} \frac{1}{k^p} \alpha + \lambda_1 \lambda_2 k$$

Considering k as a real variable we note that $E_{l=0}(k)$ is a convex function. Taking a derivative in k gives

$$\frac{dE_{l=0}}{dk} = \frac{-2p}{p+1} \left(\frac{L}{2k} \right)^{p+1} \alpha + \lambda_1 \lambda_2.$$

If $\lambda_2 \geq \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p}$ then $\frac{dE_{l=0}}{dk} \geq 0$ for $k > \bar{k}$.

If $\lambda_2 < \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p}$ then the point where the minimum is reached

$$\bar{k}_{l=0}^* = \frac{L}{2} \left(\frac{(p+1)\lambda_1\lambda_2}{2p\alpha} \right)^{-\frac{1}{p+1}}$$

satisfies $\bar{k}_{l=0}^* > \bar{k}$ and thus belongs to the range considered. If $\bar{k}_{l=0}^*$ is an integer then it is the minimizer of the energy, otherwise the minimizer is in the set $\{\lfloor \bar{k}_{l=0}^* \rfloor, \lfloor \bar{k}_{l=0}^* \rfloor + 1\}$. In all cases let us denote by $k_{l=0}^*$ the minimizer of the energy: $k_{l=0}^* = \arg \min_{k=\lfloor \bar{k}_{l=0}^* \rfloor, \lfloor \bar{k}_{l=0}^* \rfloor + 1} E_{l=0}(k)$.

We note that there is a special case that $k_{l=0}^* < \bar{k}$. In that case the minimizer of the energy with exactly $k_{l=0}^* + 1$ components will be the one considered in the analysis of the $1 \leq k \leq \bar{k}$ case, and thus will have segments of positive length l^* given by formula (A.3).

To summarize, the optimal number of components will be

$$(A.5) \quad \begin{cases} 1 & \text{if } \lambda_2 \geq \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p} \\ \lfloor \bar{k} \rfloor + 1 & \text{if } \lambda_2 < \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p}, \text{ and } k_{l=0}^* < \bar{k} \\ k_{l=0}^* + 1 & \text{if } \lambda_2 < \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p}, \text{ and } k_{l=0}^* \geq \bar{k}. \end{cases}$$

In the first case, there is just one single connected component. In the second case there are $\lfloor \bar{k} \rfloor + 1$ components, each with equal positive length. We note that by Lemma A.1 there exists a configuration with the same energy where one of these components has positive length, while the rest have zero length. The third case is that each of the components has length zero. We point out that if \bar{k} is integer-valued and $\lambda_2 < \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p}$, then the minimizer will have $k_{l=0}^* + 1$ components.

We can now derive conclusions to the structure of minimizers if $L \gg 1$. From above we conclude that the minimizer will have one component (and be a continuous line) if $\lambda_2 \geq \frac{2p}{p+1} \left(\frac{\lambda_1}{\alpha} \right)^{1/p}$, and break up into at least $\lfloor \bar{k}_{l=0}^* \rfloor + 1$ components otherwise. Rearranging, the condition also provides the critical density at which topological changes (gaps) in minimizers occur:

$$(A.6) \quad \alpha^* = \left(\frac{2p}{p+1} \right)^p \frac{\lambda_1}{\lambda_2^p}.$$

Finally we note that the typical gap length is $L/(\bar{k}_{l=0}^*)$ that is

$$(A.7) \quad L^* = 2 \left(\frac{(p+1)\lambda_1\lambda_2}{2p\alpha} \right)^{\frac{1}{p+1}}.$$

REFERENCES

- [1] E. ARIAS-CASTRO, D. L. DONOHO, AND X. HUO, *Adaptive multiscale detection of filamentary structures in a background of uniform random points*, Ann. Statist., 34 (2006), pp. 326–349.
- [2] G. BIAU AND A. FISCHER, *Parameter selection for principal curves*, Information Theory, IEEE Transactions on, 58 (2012), pp. 1924–1939.
- [3] S. BOYD, N. PARIKH, E. CHU, B. PELEATO, AND J. ECKSTEIN, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Found. Trends Mach. Learn., 3 (2011), pp. 1–122.
- [4] T. BRODERICK, B. KULIS, AND M. JORDAN, *Mad-bayes: Map-based asymptotic derivations from bayes*, in Proceedings of The 30th International Conference on Machine Learning, 2013, pp. 226–234.
- [5] G. BUTTAZZO, E. OUDET, AND E. STEPANOV, *Optimal transportation problems with free dirichlet regions*, in Variational Methods for Discontinuous Structures, G. dal Maso and F. Tomarelli, eds., vol. 51 of Progress in Nonlinear Differential Equations and Their Applications, Birkhäuser Basel, 2002, pp. 41–65.
- [6] G. BUTTAZZO AND E. STEPANOV, *Optimal transportation networks as free dirichlet regions for the monge-kantorovich problem*, Annali della Scuola Normale Superiore di Pisa - Classe di Scienze, 2 (2003), pp. 631–678.
- [7] F. CHAZAL AND J. SUN, *Gromov-hausdorff approximation of filament structure using reeb-type graph*, in Proceedings of the Thirtieth Annual Symposium on Computational Geometry, SOCG’14, New York, NY, USA, 2014, ACM, pp. 491:491–491:500.
- [8] Y.-C. CHEN, C. R. GENOVESE, AND L. WASSERMAN, *Asymptotic theory for density ridges*, Ann. Statist., 43 (2015), pp. 1896–1928.
- [9] Y.-C. CHEN, S. HO, P. E. FREEMAN, C. R. GENOVESE, AND L. WASSERMAN, *Cosmic web reconstruction through density ridges: method and algorithm*, Monthly Notices of the Royal Astronomical Society, 454 (2015), pp. 1140–1156.
- [10] R. R. COIFMAN AND S. LAFON, *Diffusion maps*, Applied and Computational Harmonic Analysis, 21 (2006), pp. 5 – 30. Special Issue: Diffusion Maps and Wavelets.
- [11] D. COMANICIU AND P. MEER, *Mean shift: A robust approach toward feature space analysis*, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24 (2002), pp. 603–619.
- [12] P. DELICADO, *Another look at principal curves and surfaces*, J. Multivariate Anal., 77 (2001), pp. 84–116.

- [13] C. DELLACHERIE AND P. MEYER, *A Probabilities and Potential*, North-Holland Mathematics Studies, Elsevier Science, 1979.
- [14] C. J. DSILVA, B. LIM, H. LU, A. SINGER, I. G. KEVREKIDIS, AND S. Y. SHVARTSMAN, *Temporal ordering and registration of images in studies of developmental dynamics*, *Development*, 142 (2015), pp. 1717–1724.
- [15] T. DUCHAMP AND W. STUETZLE, *Geometric properties of principal curves in the plane*, in *Robust statistics, data analysis, and computer intensive methods* (Schloss Thurnau, 1994), vol. 109 of *Lecture Notes in Statist.*, Springer, New York, 1996, pp. 135–152.
- [16] D. EBERLY, *Ridges in image and data analysis*, vol. 7, Springer Science & Business Media, 1996.
- [17] E. ESSER, *Applications of lagrangian-based alternating direction methods and connections to split bregman*, *CAM Report*, 31 (2009).
- [18] C. FEUERSÄNGER AND M. GRIEBEL, *Principal manifold learning by sparse grids*, *Computing*, 85 (2009), pp. 267–299.
- [19] X. GE, I. I. SAFA, M. BELKIN, AND Y. WANG, *Data skeletonization via reeb graphs*, in *Advances in Neural Information Processing Systems 24*, Curran Associates, Inc., 2011, pp. 837–845.
- [20] C. R. GENOVESE, M. PERONE-PACIFICO, I. VERDINELLI, AND L. WASSERMAN, *Nonparametric ridge estimation*, *Ann. Statist.*, 42 (2014), pp. 1511–1545.
- [21] S. GERBER AND R. WHITAKER, *Regularization-free principal curve estimation*, *The Journal of Machine Learning Research*, 14 (2013), pp. 1285–1302.
- [22] T. GOLDSTEIN AND S. OSHER, *The split Bregman method for L1-regularized problems*, *SIAM J. Imaging Sci.*, 2 (2009), pp. 323–343.
- [23] T. HASTIE AND W. STUETZLE, *Principal curves*, *J. Amer. Statist. Assoc.*, 84 (1989), pp. 502–516.
- [24] R. O. KARLSTROM AND D. A. KANE, *A flipbook of zebrafish embryogenesis*, *Development*, 123 (1996), pp. 461–462.
- [25] B. KEGL, A. KRZYZAK, T. LINDER, AND K. ZEGER, *Learning and design of principal curves*, *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22 (2000), pp. 281–297.
- [26] B. KULIS AND M. JORDAN, *Revisiting k-means: New algorithms via bayesian nonparametrics*, in *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, J. Langford and J. Pineau, eds., ICML '12, New York, NY, USA, July 2012, Omnipress, pp. 513–520.
- [27] X. Y. LU AND D. SLEPČEV, *Average-distance problem for parameterized curves*, *ESAIM: COCV*, (2015).
- [28] X. Y. LU AND D. SLEPČEV, *Properties of minimizers of average-distance problem via discrete approximation of measures*, *SIAM Journal on Mathematical Analysis*, 45 (2013), pp. 3114–3131.
- [29] C. MANTEGAZZA, A. C. MENNUCCI, ET AL., *Hamilton-Jacobi equations and distance functions on riemannian manifolds*, *Applied Mathematics and Optimization*, 47 (2003), pp. 1–26.
- [30] S. OSHER, M. BURGER, D. GOLDFARB, J. XU, AND W. YIN, *An iterative regularization method for total variation-based image restoration*, *Multiscale Modeling & Simulation*, 4 (2005), pp. 460–489.
- [31] U. OZERTEM AND D. ERDOGMUS, *Locally defined principal curves and surfaces*, *The Journal of Machine Learning Research*, 12 (2011), pp. 1249–1286.
- [32] S. PULKKINEN, *Ridge-based method for finding curvilinear structures from noisy data*, *Computational Statistics & Data Analysis*, 82 (2015), pp. 89 – 109.
- [33] A. SINGER AND H.-T. WU, *Vector diffusion maps and the connection Laplacian*, *Comm. Pure Appl. Math.*, 65 (2012), pp. 1067–1144.
- [34] G. SINGH, F. MEMOLI, AND G. CARLSSON, *Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition*, in *Eurographics Symposium on Point-Based Graphics*, 2007, pp. 91–100.
- [35] D. SLEPČEV, *Counterexample to regularity in average-distance problem*, *Annales de l’Institut Henri Poincaré (C) Non Linear Analysis*, 31 (2014), pp. 169 – 184.
- [36] A. J. SMOLA, S. MIKA, B. SCHÖLKOPF, AND R. C. WILLIAMSON, *Regularized principal manifolds*, *J. Mach. Learn. Res.*, 1 (2001), pp. 179–209.
- [37] R. TIBSHIRANI, *Principal curves revisited*, *Stat. Comput.*, 2 (1992), pp. 182–190.
- [38] R. TIBSHIRANI, M. SAUNDERS, S. ROSSET, J. ZHU, AND K. KNIGHT, *Sparsity and smoothness via the fused lasso*, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (2005), pp. 91–108.
- [39] J. VERBEEK, N. VLASSIS, AND B. KROSE, *A k-segments algorithm for finding principal curves*, *Pattern Recognition Letters*, 23 (2002), pp. 1009 – 1017.
- [40] H. WANG AND T. C. LEE, *Automatic parameter selection for a k-segments algorithm for computing principal curves*, *Pattern recognition letters*, 27 (2006), pp. 1142–1150.